MACHINE LEARNING APPROACHES TO PREDICT CROP YIELD: A STUDY OF MODEL PERFORMANCE ON SEASONAL WEATHER AND LOCATION DATA IN NEPAL

In partial fulfillment of

FATHERS LOCKE AND STILLER RESEARCH AWARDS

(LSRA)

MINI THESIS

LSRA CYCLE 7

Submitted to the

Department of Research

St. Xavier's College

Maitighar, Kathmandu, Nepal

By

AAYUSH SHRESTHA

Department of Computer Science

ST. XAVIER'S COLLEGE

KATHMANDU, NEPAL

CERTIFICATE OF APPROVAL

The head of research hereby makes known that ***Aayush Shrestha, B.Sc. CSIT 3rd year, Department of Computer Science, 021BSCIT004*** has been conferred the **Locke And Stiller Research Award (Lsra) Cycle 7 for his** research work embodied in this mini-thesis entitled **"Machine Learning Approaches To Predict Crop Yield: A Study Of Model Performance On Seasonal Weather And Location Data In Nepal"** under the supervision and guidance of ***Mr. Ramesh Shahi, Department of Computer Science*** for the full period prescribed by the LSRA-research office of St. Xavier's College, Maitighar, Kathmandu, Nepal.

Department of Computer Science
Place: St. Xavier's College, Kathmandu
Date:

…………………
Mr. Niraj Nakarmi

RECOMMENDATION FOR APPROVAL

This is to certify that ***Aayush Shrestha, B.Sc. CSIT 3rd year, Department of Computer Science, 021BSCIT004*** has carried out the **FATHERS LOCKE AND STILLER RESEARCH AWARD (LSRA)** research work embodied in this mini-thesis under the supervision and guidance of ***Mr. Ramesh Shahi, Department of Computer Science*** for the full period prescribed by the LSRA-research office of St. Xavier's College, Maitighar, Kathmandu, Nepal. We approve this mini-thesis entitled **"Machine Learning Approaches To Predict Crop Yield: A Study Of Model Performance On Seasonal Weather And Location Data In Nepal"** for submission for the **FATHERS LOCKE AND STILLER RESEARCH AWARD (LSRA) CYCLE 7.**

Department of Computer Science
Place: St. Xavier's College, Kathmandu
Date:

…………	…………
Mr. Ganesh Yogi	Er. Anuj Shrestha
Head of Computer Science Department	Department Research Coordinator for LSRA
St. Xavier's College	St. Xavier's College

CERTIFICATE OF RECOMMENDATION

This is to certify that this mini-thesis entitled **"*Machine Learning Approaches To Predict Crop Yield: A Study Of Model Performance On Seasonal Weather And Location Data In Nepal*"** submitted by ***Aayush Shrestha, B.Sc. CSIT 3rd year, Department of Computer Science, 021BSCIT004*** for the **FATHERS LOCKE AND STILLER RESEARCH AWARD (LSRA) CYCLE 7,** is a record of research work done by the candidate from **May 2024** to **February 2025** at the Department of ***Computer Science*,** St. Xavier's College, Maitighar, Kathmandu, Nepal, under my supervision and guidance and has not previously been submitted for any degree, diploma or project work.

I recommend this mini-thesis for submission.

Department of Computer Science
Place: St. Xavier's College
Date:

………………………
Mr. Ramesh Shahi

DECLARATION

I, ***Aayush Shrestha, B.Sc. CSIT 3rd year, Department of Computer Science, 021BSCIT004*** hereby declare that the work presented in the mini thesis entitled **"*Machine Learning Approaches To Predict Crop Yield: A Study Of Model Performance On Seasonal Weather And Location Data In Nepal*"** is a record of independent research work done by me from **May 2024 To February 2025** at ***Department of Computer Science*,** St. Xavier's College, Maitighar, Kathmandu, Nepal, under the supervision and guidance of ***Mr. Ramesh Shahi, Department of Computer Science***, St. Xavier's College, Maitighar, Kathmandu, Nepal. This work or part of it has not previously been submitted for any degree, diploma or project work.

I submit this mini-thesis for the **LOCKE AND STILLER RESEARCH AWARD (LSRA) CYCLE 7.**

Department of Computer Science
Place: St. Xavier's College, Kathmandu
Date:

………………

Aayush Shrestha

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, **Mr. Ramesh Shahi**, for his guidance and support throughout this research. I am also sincerely thankful to **Er. Anuj Shrestha**, the Department Research Coordinator, for his valuable insights and encouragement in shaping my research.

I extend my appreciation to the **Research Department** for providing me with this opportunity to learn and grow through the **Fathers Locke and Stiller Research Awards (LSRA)**. This journey has been instrumental in helping me understand the disciplines of research and refine my academic skills.

Lastly, I am profoundly grateful to my **friends and family** for their unwavering support and encouragement throughout this process. Their motivation and belief in me have been invaluable.

……………………

AAYUSH SHRESTHA
DEPARTMENT OF COMPUTER SCIENCE

LISTS OF TABLES

LISTS OF FIGURES

ABBREVIATIONS

AI	Artificial Intelligence
CSV	Comma-Separated Values
EDA	Exploratory Data Analysis
GIS	Geographic Information System
GPS	Global Positioning System
IQR	Interquartile Range
IoT	Internet of Things
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
ML	Machine Learning
NDVI	Normalized Difference Vegetation Index
PPT	Precipitation
RF	Random Forest
R^2	Coefficient of Determination
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SVR	Support Vector Regression
TEMP	Temperature

ABSTRACT

Food security is a pressing global challenge, particularly in countries like Nepal, where agriculture is heavily reliant on unpredictable weather patterns. Accurate crop yield prediction is vital for informed decision-making and resource management. However, the lack of models incorporating weather and location data presents a significant knowledge gap. This study aims to evaluate and compare machine learning models for crop yield prediction, focusing on five major cereal crops in Nepal: barley, maize, millet, paddy, and wheat.

The study employed Random Forest, Support Vector Regression (SVR), and Linear Regression to predict crop yields based on a comprehensive dataset of weather and agricultural variables. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 metrics. Feature importance analysis was conducted to identify key factors influencing yield predictions.

The Random Forest model outperformed other models across all crops, with higher accuracy and robustness. Key influencing factors included rainfall, temperature, and humidity, with their seasonal variations playing a critical role. Residual analysis highlighted a conical error distribution, with smaller yields exhibiting fewer errors and larger yields showing greater variance. Outliers were identified in crops such as maize and millet, indicating opportunities for dataset refinement.

This study underscores the potential of machine learning to improve crop yield prediction in Nepal, providing actionable insights for policymakers and stakeholders. Future research should explore integrating real-time weather data and advanced neural networks for enhanced predictive accuracy.

Keywords: Crop Yield Prediction, Machine Learning, Random Forest, Weather Data, Nepal Agriculture

CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Background And Importance

Agriculture is a major factor contributing to Nepal's GDP averaging 26.1% in the last 10 years, Government of Nepal (2024), and sustains the bulk of our population. However, it is burdened by many challenges like unpredictable weather patterns, limited adoption of new technologies, land degradation, improper use of budget, and lack of support towards farmers (Gyawali & Khanal, 2021; Khanal et al., 2020).

According to Kattel & Nepal (2022) about 65% of the total cultivable land in Nepal is supported by rain-based agriculture, and only 24% of the agricultural land is irrigated, leaving the country mostly dependent on natural and timely rainfall for a good yield.

Cereal crops are vital for addressing food security in Nepal. They occupy a significant share of agricultural land and production, with paddy, maize, and wheat being the most prominent. These crops are essential for the nutrition and sustenance of the Nepalese population (Dahal et al., 2022).

1.2 Problem Statement

As more than 80% of the precipitation occurs between June and September, this rapid downpour regularly leads to flooding and landslides that destroy croplands (Malla, 2009). This arrangement leaves the country vulnerable to drastically varying degrees of crop yield depending on the characteristics of that year's monsoon.

According to Sigdel et al. (2022) The absence of predictive tools tailored to Nepal's unique agricultural landscape significantly hinders progress. Farmers face economic constraints, like high costs of technology and limited financial resources, deterring them from adopting new approaches. Furthermore, a lack of awareness and understanding of the benefits of agricultural technologies, coupled with insufficient training and education, exacerbates the issue. These challenges, combined with the absence of locally adapted predictive models, impede the effective utilization of data-driven insights for improving agricultural practices and mitigating risks (2022).

Therefore, crop yield prediction is a difficult but invaluable technique to increase agricultural productivity and reduce food insecurity (Van Klompenburg, Kassahun, & Catal, 2020).

1.3 Role Of Machine Learning

The integration of machine learning with big data computing paradigms offers promising avenues for improving crop yield forecasting. These predictive models can help address challenges such as water shortages, climate uncertainty, and soil fertility reduction, ultimately supporting smart farming practices and food security (Palanivel & Surianarayanan, 2019).

An accurate prediction model can provide foresight for the concerned authorities to balance the trade of agricultural products. An early warning system can mitigate the adverse effects that often plague farmers in our country (A. Kumar et al., 2019).

This study employs three different machine learning models to analyze complex relationships within the datasets and generate predictions. Random Forest (RF), an ensemble learning method, leverages multiple decision trees to capture non-linear relationships and improve predictive accuracy (Schonlau & Zou, 2020; Scornet et al., 2015). Support Vector Regression (SVR) excels in handling outliers and modeling non-linear relationships through kernel functions (Üstün et al., 2007; Zhang & O'Donnell, 2020). Linear Regression, while simpler, provides a baseline for comparison and effectively models linear relationships (Dongarra et al., 2011).

These models were selected for their versatility and ability to address different aspects of agricultural prediction, ensuring a comprehensive and robust analysis.

1.4. Objective Of The Study

Owing to the above facts, by utilizing historical crop yield data, weather patterns, and geographical distribution of arable land, this paper aims to construct a robust tool for the farmers and policymakers in Nepal. This research aims to develop a machine-learning model capable of addressing the unpredictability of crop yields.

The primary focus of this model will be to identify the major factors that affect crop yield and perform a comparative analysis of the various machine learning algorithms to recognize an effective model for our specific circumstances.

1.5. Scope And Contribution

The findings of this study contribute to the growing body of knowledge on the application of machine learning in agriculture, specifically in the context of Nepal. This research introduces a unique approach by focusing on the comparative performance of machine learning models—Random Forest, SVR, and Linear Regression—on predicting cereal crop yields using location-specific weather data. The study's emphasis on Nepalese agricultural challenges, such as monsoon dependency and limited irrigation, sets it apart from global research and provides region-specific insights.

By integrating historical crop yield and weather data, this research not only identifies the most effective model for Nepal’s conditions but also highlights critical factors influencing yield variability. These findings offer practical solutions to improve crop management strategies, optimize resource allocation, and support policymaking to reduce food insecurity.

Moreover, this work lays a foundation for future studies in crop yield prediction, encouraging the exploration of additional machine learning techniques, expanded datasets, and the inclusion of socioeconomic and soil data. It aims to inspire further advancements in precision agriculture, ultimately fostering sustainable farming practices and supporting economic growth in the region.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction To Crop Yield Prediction

2.1.1 Overview of Crop Yield Prediction

From Van Klompenburg et al.(Van Klompenburg et al., 2020), crop yield prediction refers to the process of estimating the quantity of crop production expected from a given area of land. This estimation depends on a variety of factors, including climatic conditions, soil characteristics, farming practices, and crop genetics. Accurate crop yield predictions are crucial for optimizing resource allocation, ensuring food security, and guiding policy decisions (Sakshi Mundhe et al., 2023).

Crop yield prediction is a critical area of research in agricultural sciences, extensively studied to address food security challenges. Studies highlight that traditional prediction methods, such as trend analysis and empirical models, often fail to account for the complex interactions between variables like weather, soil, and farming practices. As a result, inaccuracies in predictions have led to inefficiencies in resource planning and management (Fan et al., 2015; Ruchira C. Mahore & Naresh G. Gadge, 2023).

Recent research from Islam et al. (2023) and Y. Wang et al. (2017) indicates that machine learning models are increasingly adopted to overcome these limitations. These models are capable of processing large datasets and uncovering intricate, non-linear relationships, leading to improved predictive accuracy. The integration of remote sensing data and advanced computational techniques has further enhanced the applicability of these methods in diverse agricultural settings (Khaki et al., 2020).

2.1.2 Importance of Accurate Crop Yield Prediction

The importance of accurate crop yield prediction has been underscored in various studies as a means to mitigate risks in agriculture (Al-Gaadi et al., 2016; Nyéki & Neményi, 2022). Research demonstrates that predictive models enable better preparedness for challenges such as adverse weather conditions, pest outbreaks, and water scarcity. This preparedness helps stakeholders minimize economic losses and optimize resource utilization ("Crop Yield Prediction and Fertilizer Recommendation," 2020; A. Kumar et al., 2019) .

Accurate crop yield predictions enable governments, agricultural stakeholders, and farmers to make informed decisions. These predictions help mitigate the risks of food insecurity, assist in planning storage and distribution, and optimize agricultural practices (Bondre & Mahagaonkar, 2019; A. Kumar et al., 2019; Ruchira C. Mahore & Naresh G. Gadge, 2023). For developing nations like Nepal, where agriculture contributes significantly to the economy, reliable predictions are essential to address challenges posed by climate change and population growth (Ayoub Shaikh et al., 2022; Sharma et al., 2021).

Moreover, Ansarifar et al. (2021) and Renju et al. (2022) revealed that accurate yield predictions support policymaking by informing strategies for food distribution, import/export planning, and disaster relief measures. For farmers, these models serve as decision-support tools for efficient crop management, including fertilizer application and irrigation scheduling. The potential of these predictions to stabilize food systems and markets has been widely documented in recent literature.

Additionally, incorporating variables such as fertilizer use and irrigation into prediction models further enhances their accuracy. These models, as noted by Nyéki and Neményi (2022) and Pradeep et al. (2023), reduce environmental impacts by optimizing resource allocation and promoting sustainable farming practices.

2.1.3 Challenges in Nepal’s agriculture

Nepal is facing a lot of challenges in the sector of agriculture and advancements in farming practices. According to Gyawali & Khanal (2021), agricultural growth in Nepal has been low and vulnerable because of technological, financial and resource constraints. Poverty, outmigration of labour, increase in population and decreasing cultivatable land has further impeded the progress of agricultural development (Gyawali & Khanal, 2021).

Limited access and availability of fetilizers, water irrigation and fair market prices of agricultural products prose a significant threat for the farmers in Nepal (Krupnik et al., 2021; Nepal et al., 2024). According to Krupnik et al. (2021), farmers practice closed farming practices with limited external inputs, leading to a low but stable yield of crops. They also noted that the hills face soil erosion and land degradation, while the flatlands or terai faces issues like reduced soil fertility and lack of irrigation. This is further corroborated by Malla (2009), where he expressed that climate variability and extreme weather plagues the entire country.

2.2 Role Of Weather And Location Data In Crop Yield Prediction

2.2.1 Weather Factors in Agriculture

Numerous studies emphasize the significant influence of weather parameters on crop yield. Variables such as temperature, rainfall, and humidity have been identified as critical determinants in agricultural productivity. Research shows that deviations in rainfall patterns, particularly during sowing or harvesting periods, can drastically affect crop yields (Archana & Kumar, 2023; Ghimire, 2019; Ngoune Liliane & Shelton Charles, 2020). Similarly, extreme temperatures, whether heatwaves or frost, have been linked to reduced crop growth and yield potential, as documented in several studies (Kong et al., 2023; Malla, 2009; Wan et al., 2022; Y. Wang et al., 2020).

As reported by (Van Klompenburg et al., 2020), the factors used for training the model depended on the available dataset. They also added that the success of a model was not proportional to the number of parameters, instead it was the converse in most scenarios. The presence of numerous factors required careful tuning of the model, which when done improperly resulted in poor performance of the models.

From the study by (Rashid et al., 2021), it was seen that climatic data and historical crop yield data were utilized in over 80% of the paper. The most used features were vegetation index and satellite data, meanwhile, the least used features were soil property, irrigation information, and cropland information like soil pH. They also concluded that DNN models were more successful than simple Machine learning algorithms, but only when the dataset was significantly large.

Recent advancements have focused on incorporating real-time weather data into predictive models. For instance, satellite-based monitoring systems and weather stations provide high-resolution datasets that enhance the temporal accuracy of yield forecasts. Studies further highlight the importance of combining historical and real-time weather data to develop robust models capable of adapting to unpredictable climatic shifts (Cai et al., 2019; Lohitha Reddy & Siva Kumar, 2023; Schwalbert et al., 2020; Van Klompenburg et al., 2020).

Further studies have explored the temporal and spatial variation of weather patterns to better understand their impact on crop yields. For example, research has shown that seasonal weather variations, such as monsoon timing and intensity, can significantly alter the productivity of crops like maize and rice (Fernandez-Beltran et al., 2021; Kuradusenge et al., 2023).

Models that integrate historical weather data have been developed to identify trends and patterns, allowing for predictions based on expected seasonal anomalies. Moreover, the increasing variability of climate due to global warming has led to a greater focus on climate-resilient crop varieties and adaptive farming techniques. Lohitha Reddy & Siva Kumar (2023), Van Klompenburg et al. (2020), and Crane-Droesch (2018) have noted that yield forecasts can be more precise when weather data is combined with crop-specific growth stages, accounting for how different crops respond to specific climatic conditions during key development phases.

2.2.2 Location factors in Agriculture

The role of geographical location is another recurring theme in crop yield prediction studies. Variations in soil types, elevation, and access to water resources introduce unique challenges in different regions. Research has consistently shown that location-specific factors create distinct agro-ecological zones, each requiring tailored prediction models (Adamopoulos & Restuccia, 2018; Ngoune Liliane & Shelton Charles, 2020). For instance, crops grown in mountainous regions may face entirely different constraints compared to those in plains, even under similar climatic conditions (C. Yang et al., 1998).

Researchers like Adamopoulos & Restuccia, (2018) and Tiwari & Shukla, (2019) also point out that the resolution of location data significantly impacts model accuracy. High-resolution spatial data, such as those derived from geographic information systems (GIS), have been found to improve predictions by accounting for micro-level variations in environmental conditions. However, challenges such as limited data availability in remote or under-researched areas remain prominent in the field (Nepal et al., 2024).

2.3 Advancements In Machine Learning For Crop Yield Prediction

2.3.1 Traditional vs Machine Learning Methods

Historically, crop yield prediction relied on statistical methods such as linear regression, time series analysis, and expert-based models. These traditional techniques, though useful, often struggled to capture the complex, non-linear relationships inherent in agricultural data (Elavarasan & Vincent, 2020; Fan et al., 2015; Palanivel & Surianarayanan, 2019; Van Klompenburg et al., 2020; You et al., 2017). As a result, they tended to offer less accurate predictions when faced with diverse variables such as varying soil conditions, climate change, and new farming practices. Researchers have highlighted that while traditional methods could handle simpler datasets, they often fell short in environments with high data variability (Fan et al., 2015; Ismail & Yusof, 2022; Rale et al., 2019; Renju et al., 2022; Sharma et al., 2021).

The introduction of machine learning (ML) has significantly changed the landscape of crop yield prediction. ML algorithms, particularly supervised learning methods like Random Forests, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), offer the ability to model complex relationships between input variables and yield outcomes. Several studies have demonstrated the superior performance of ML models in terms of predictive accuracy when compared to traditional approaches. For instance, ML models are better equipped to handle large, multidimensional datasets that include weather, soil properties, and farming techniques (Araǘjo et al., 2023; Ismail & Yusof, 2022; Rashid et al., 2021)

2.3.2 Machine learning models used in crop yield prediction

In recent years, deep learning models, particularly Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN), have shown significant potential in crop yield prediction. These models, part of the broader family of artificial neural networks (ANN), require large and complex datasets for training, making them suitable for handling high-dimensional data sources such as satellite imagery, remote sensing data, and multi-year weather patterns (Fernandez-Beltran et al., 2021; Khaki et al., 2020; Rashid et al., 2021; Van Klompenburg et al., 2020).

CNNs, in particular, are often used for spatial data analysis, making them effective for extracting features from images or geographical data related to crop fields. As Khaki et al.(2020) and Tanabe et al. (2023) have demonstrated that CNNs can capture fine-grained spatial patterns that contribute to accurate predictions of crop yield.

Furthermore, according to Kuwata & Shibasaki (2016) and Rashid et al. (2021) DNNs, which consist of multiple hidden layers, are adept at modeling non-linear relationships and capturing intricate patterns across diverse input features. They have found that DNNs perform particularly well when working with complex, high-volume datasets, such as time-series weather data combined with location-specific variables. Their ability to learn hierarchical features allows them to handle a wide range of inputs, including historical weather data, soil conditions, and even socio-economic factors, which traditional models struggle to incorporate (Rashid et al., 2021; Van Klompenburg et al., 2020).

The use of DNNs in crop yield prediction has grown, with many studies reporting high accuracy rates compared to simpler models. However, the complexity of training these models and the need for vast amounts of data pose make them for impractical for application in regions with limited data availability (Kim et al., 2019; Kuwata & Shibasaki, 2016).

According to Araújo et al. (2023), the most used machine learning algorithms to predict crop yield in their review were Random Forest(RF) at 19.2%, followed by Support Vector Machine(SVM) at 15.9%, Gradient Boosted Tree(GBT) at 8.3% and CNN with a 7.3% of the overall composition.

Random Forest (RF) has become one of the most widely used machine learning algorithms for crop yield prediction due to its robustness and versatility (Rashid et al., 2021; Van Klompenburg et al., 2020). RF is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy (Schonlau & Zou, 2020).

Saeed et al. (2017) have shown that RF excels in handling both continuous and categorical data and is less prone to overfitting, making it ideal for complex agricultural datasets.

Its ability to provide feature importance rankings is another key strength, allowing researchers to identify the most influential factors affecting yield prediction, such as rainfall or soil nutrients. In many studies, RF has outperformed other models in terms of prediction accuracy and generalizability, especially when applied to diverse crops and regions (Schonlau & Zou, 2020).

Support Vector Regression (SVR), a type of Support Vector Machine (SVM), is another powerful model that has been widely used in agricultural prediction tasks. SVR is particularly known for its ability to model non-linear relationships by using kernel functions to map input data into higher-dimensional spaces. Research has shown that SVR performs well even in situations where the dataset is small or noisy, as it focuses on finding the optimal hyperplane that minimizes error while maintaining model generalization (Mallikarjuna Rao et al., 2022; Priyadharshini et al., 2022).

Further as reported by Mallikarjuna Rao et al. (2022) and Priyadharshini et al. (2022), SVR has been found to be particularly effective for predicting crop yields when only limited historical data is available. However, the model's performance can degrade when working with highly complex datasets unless adequate preprocessing, feature engineering and hyperparameter tuning is employed.

2.3.3 Feature Importance in ML Models

An important aspect of machine learning in crop yield prediction is the identification of feature importance. By determining which variables contribute most to yield predictions, researchers can gain valuable insights into the factors driving agricultural outcomes. Feature importance analysis is often integrated into models like Random Forest, which inherently provides a ranking of variables based on their contribution to predictive accuracy. Rashid et al. (2021) and Van Klompenburg et al. (2020) have pointed out that weather variables such as rainfall and temperature are often among the most influential factors, followed by soil characteristics and crop-specific factors.

This ability to prioritize features not only aids in improving model performance but also offers practical insights for farmers and policymakers. By focusing on the most impactful variables, interventions can be better targeted, such as optimizing irrigation practices based on predicted rainfall or choosing crop varieties suited to the expected temperature range. However, challenges remain in accounting for the dynamic nature of these features over time, which requires continuous model refinement (Rashid et al., 2021; Van Klompenburg et al., 2020).

2.4 Insights From Previous Studies

2.4.1 Global Perspectives

Globally, significant research has been conducted to evaluate crop yield prediction models under diverse agricultural and climatic conditions. Studies from around the globe have often focused on integrating high-resolution weather, satellite imagery and soil datasets with advanced machine learning techniques (Van Klompenburg et al., 2020).

For instance, Kim et al. (2019), Kuwata & Shibasaki (2016), and Y. Wang et al. (2020) all in the United States has demonstrated the efficacy of Random Forest and Deep Neural Network models in predicting yields for staple crops such as corn and soybeans. Following suit, a study conducted in Brazil by Schwalbert et al. (2020) led to similar findings for the prediction of yield for soybeans.

Similarly, Cai et al. (2019) demonstrated that using satellite and climate data, were effective at predicting crop yield using machine learning models like Random Forest, Neural Networks and Support Vector Machines in Australia.

Furthermore, European studies have shown promising results using machine learning models for analyzing weather data to forecast wheat and other crop production (Cantelaube & Terres, 2005; Paudel et al., 2022).

In developing regions, researchers have highlighted the potential of machine learning models to address food security challenges. Studies in sub-Saharan Africa, for example, have focused on using models such as SVR, Deep learning, and ensemble learning methods. These models have been applied successfully to predict yields of drought-resistant crops like millet and sorghum, showing that models can be effective when adapted to local contexts (Aworka et al., 2022; Petersen, 2018).

2.4.2 Studies Focused on Nepal

In Nepal, research on crop yield prediction is still in its nascent stages, with increasing interest in leveraging machine learning to address the country's unique agricultural challenges. Studies have shown that Nepal's diverse agro-climatic zones, ranging from lowland Terai to high-altitude Himalayan regions, require location-specific models tailored to varying environmental and socio-economic conditions (Nepal et al., 2024; Sigdel et al., 2022).

Recent studies have also explored the integration of remote sensing data with traditional agricultural statistics to improve prediction accuracy. For instance, satellite-derived vegetation indices, when combined with ground-level data, have been used to predict paddy yields in the Terai region with a high degree of accuracy (Fernandez-Beltran et al., 2021; Sigdel et al., 2022).
Dahal et al. (2022) has identified weather variables such as monsoon rainfall and temperature as key determinants for major cereal crops like rice, wheat and paddy. Similarly, other models have been developed that are tailored to predict specific crop yields like rice using meteorological data and climate data (Islam et al., 2023).

2.5 Challenges And Limitations

2.5.1 Data Availability and Quality

One of the most significant challenges in crop yield prediction using machine learning is the availability and quality of data. Many datasets, especially in developing countries, are incomplete, outdated, or lack consistency. For example, weather data from ground-based stations is often sparse or limited to certain regions, leaving large geographical areas unrepresented (Rashid et al., 2021; Van Klompenburg et al., 2020). Additionally, agricultural data such as crop yields, soil properties, and farming practices may be collected infrequently or rely on estimations, leading to inaccuracies.

Studies have also highlighted the issues of data granularity and resolution. High-resolution data, such as hourly weather measurements or field-level crop statistics, is often required for precise model training but is rarely available. Instead, researchers must often work with aggregated datasets, such as district- or state-level averages, which can obscure important variations (Sharma et al., 2021). Furthermore, challenges in integrating data from multiple sources, such as combining satellite imagery with ground-level data, can introduce discrepancies and require extensive preprocessing (Schwalbert et al., 2020; Tiwari & Shukla, 2019).

2.5.2 Model Limitations

Despite their advantages, machine learning models also face several limitations when applied to crop yield prediction. Overfitting remains a persistent challenge, particularly when models are trained on small datasets or those with high variability. Models such as Deep Neural Networks and Convolutional Neural Networks require vast amounts of labeled data for optimal performance, which may not always be available (Khaki et al., 2020; Sakib et al., 2018).

Interpretability is another critical issue. Many machine learning models, particularly deep learning approaches, are often referred to as "black boxes" due to their lack of transparency in explaining how predictions are made. This lack of explainability can limit their adoption where we require clear and actionable insights (Rudin, 2019). Moreover, computational complexity can pose a barrier, as training sophisticated models requires significant processing power, which may not be feasible in regions with limited technical infrastructure (Sakib et al., 2018; Van Klompenburg et al., 2020).

2.5.3 Issues in Nepal

In Nepal, the challenges associated with data availability and model limitations are further compounded by the country's unique geographical and socio-economic conditions. The lack of comprehensive, high-quality datasets is a major hurdle (Fernandez-Beltran et al., 2021).

Agricultural data, including crop yield and production statistics, is often aggregated at the district level, which may not accurately reflect field-level variability (Agriculture Statistics | Publication Category | Ministry of Agriculture and Livestock Development, n.d.).

Additionally, Nepal's diverse geography and climatic zones make it difficult to develop a single predictive model applicable across the country. The variability in rainfall patterns, temperature ranges, and soil types requires localized models, which are resource-intensive to develop and maintain (Dahal et al., 2022; Nepal et al., 2024).

Furthermore, the lack of technical expertise and infrastructure in many parts of the country hampers the adoption of advanced machine learning techniques. Researchers also face challenges in integrating traditional agricultural practices and indigenous knowledge into machine learning frameworks, which could enhance model relevance and acceptance (Gyawali & Khanal, 2021; Khanal et al., 2020; Sigdel et al., 2022).

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Design

The research design for this study is centered on building machine learning models to predict crop yields in Nepal using weather and crop data from 1979 to 2022. The study integrates district-level crop statistics with weather variables such as temperature, rainfall, and humidity to assess the impact of climatic factors on agricultural productivity.

The methodology consists of six distinct phases:

1. **Data Collection**: Obtaining weather data from NASA's climate portal and crop yield data from Nepal's government publications.
2. **Data Processing and Feature Engineering**: Addressing inconsistencies in district-level data, handling missing values, and creating seasonal weather features.
3. **Dataset Consolidation**: Merging weather and crop yield data into a unified dataset for analysis.
4. **Model Implementation**: Training and optimizing Random Forest, Support Vector Regression (SVR), and Linear Regression models.
5. **Model Evaluation**: Comparing the models' performance using metrics such as RMSE, MAE, and R^2 score.
6. **Model Result Visualization**: visualizing the results obtained from the model using graphics like *actual vs predicted yields*, *feature importance*, and *residual plots*.

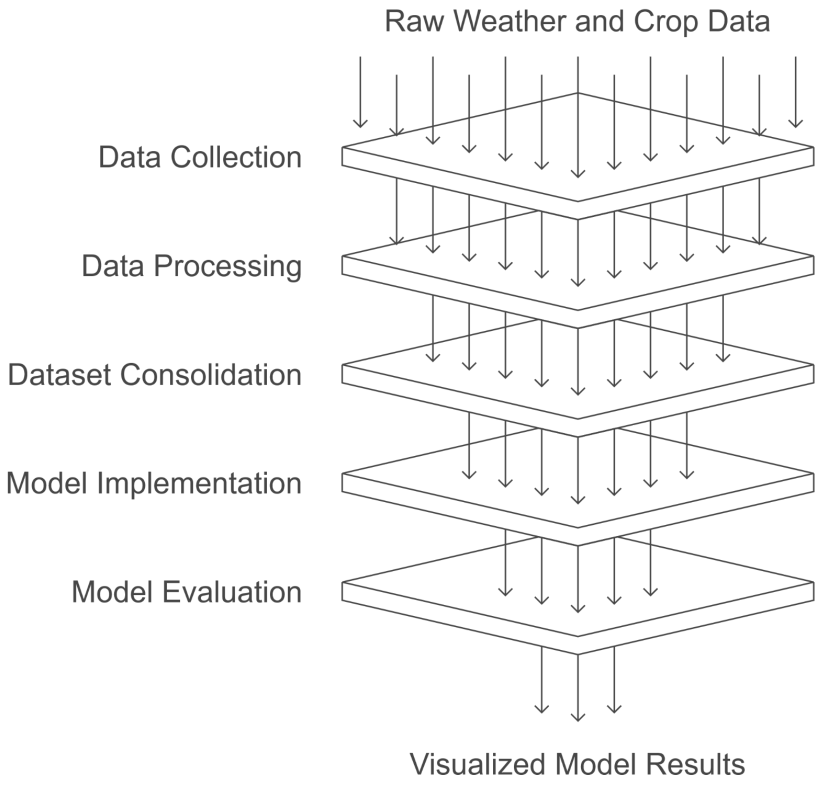


Figure 1. *Six distinct phases of the research methodology*

The choice of models reflects their ability to handle regression tasks with varying levels of complexity, from ensemble learning in Random Forest to non-linear relationships in SVR. Hyperparameter tuning using grid search ensured optimal model performance.

Challenges, such as changing district boundaries and missing data, were addressed using interpolation techniques and neighboring district averages. Seasonal trends in weather data were derived to capture the temporal patterns influencing crop growth.

The following sections provide a detailed explanation of each phase of the methodology.

3.2 Data Collection

The data collection phase involved gathering two key datasets essential for modeling crop yields: weather data and crop statistics. This section outlines the data sources, formats, and key variables collected, along with any preprocessing steps undertaken during acquisition.

3.2.1 Weather Data Collection

The weather data was collected from the weather data was sourced from Open Data Nepal, who consolidated weather data from NASA Langley Research Center (LaRC) (*District Wise Climate Data for Nepal - District Wise Monthly Climate Data for Nepal - Open Data Nepal*, n.d.). The data spans from 1981-01-01 to 2019-12-31 and includes the following features:

1. PRECTOT: MERRA2 1/2x1/2 Precipitation (mm day-1)
2. PS: MERRA2 1/2x1/2 Surface Pressure (kPa)
3. QV2M: MERRA2 1/2x1/2 Specific Humidity at 2 Meters (g/kg)
4. RH2M: MERRA2 1/2x1/2 Relative Humidity at 2 Meters (%)
5. T2M: MERRA2 1/2x1/2 Temperature at 2 Meters (C)
6. T2MWET: MERRA2 1/2x1/2 Wet Bulb Temperature at 2 Meters (C)
7. T2M_MAX: MERRA2 1/2x1/2 Maximum Temperature at 2 Meters (C)
8. T2M_MIN: MERRA2 1/2x1/2 Minimum Temperature at 2 Meters (C)
9. T2M_RANGE: MERRA2 1/2x1/2 Temperature Range at 2 Meters (C)
10. TS: MERRA2 1/2x1/2 Earth Skin Temperature (C)
11. WS10M: MERRA2 1/2x1/2 Wind Speed at 10 Meters (m/s)
12. WS10M_MAX: MERRA2 1/2x1/2 Maximum Wind Speed at 10 Meters (m/s)
13. WS10M_MIN: MERRA2 1/2x1/2 Minimum Wind Speed at 10 Meters (m/s)
14. WS10M_RANGE: MERRA2 1/2x1/2 Wind Speed Range at 10 Meters (m/s)
15. WS50M: MERRA2 1/2x1/2 Wind Speed at 50 Meters (m/s)
16. WS50M_MAX: MERRA2 1/2x1/2 Maximum Wind Speed at 50 Meters (m/s)

17. WS50M_MIN: MERRA2 1/2x1/2 Minimum Wind Speed at 50 Meters (m/s)
18. WS50M_RANGE: MERRA2 1/2x1/2 Wind Speed Range at 50 Meters (m/s)

These variables were selected for their relevance to crop growth and yield. Data was extracted on a district-level basis, ensuring compatibility with the spatial granularity of the crop yield dataset, which contained the crop yield on a per district basis.

Table 1. *Sample snapshot of the weather dataset*

id	YEAR	MONTH	DISTRICT	PRECTOT	PS	QV2M
1	1981	1	Arghakhanchi	67.31	93.78	5.28
2	1981	2	Arghakhanchi	3.37	93.52	5.13
3	1981	3	Arghakhanchi	26.02	93.4	5.91
4	1981	4	Arghakhanchi	46.15	93.03	6.52
5	1981	5	Arghakhanchi	69.45	92.75	9.95
6	1981	6	Arghakhanchi	110.77	92.35	12.04
7	1981	7	Arghakhanchi	620.05	92.43	19.43
8	1981	8	Arghakhanchi	309.99	92.46	19.55
9	1981	9	Arghakhanchi	184.98	93.02	17.22
10	1981	10	Arghakhanchi	11.72	93.34	10.56
11	1981	11	Arghakhanchi	36.66	93.56	7.72
12	1981	12	Arghakhanchi	1.37	93.9	4.83

Note. This table contains the first 12 rows of the dataset, with the first 7 columns. The actual database contains more columns.

PRECTOT stands for Total Precipitation, PS stands for Surface Pressure, and QV2M stands for Specific Humidity at 2 Meters.

3.2.2 Crop Yield Data Collection

Crop statistics were collected from the agricultural statistics published by the ministry of Agriculture and Livestock Development (*Agriculture Statistics* |

Publication Category | Ministry of Agriculture and Livestock Development, n.d.). The data spans from 1979-1 to 2022-12.

These publications contained district wise information about the five major cereal crops (Barley, Maize, Millet, Paddy and Wheat). It included the Area in hectares, Production in metric tons, and Yield in metric tons per hectare.

The data was available in a PDF format. To extract the data into a CSV (Comma Separated Variables) file, OCR (Optical Character Recognition) was deployed to extract the tabular dataset. The extracted data was then cleaned and aligned to ensure consistency. A sample of the published dataset is shown is the following section.

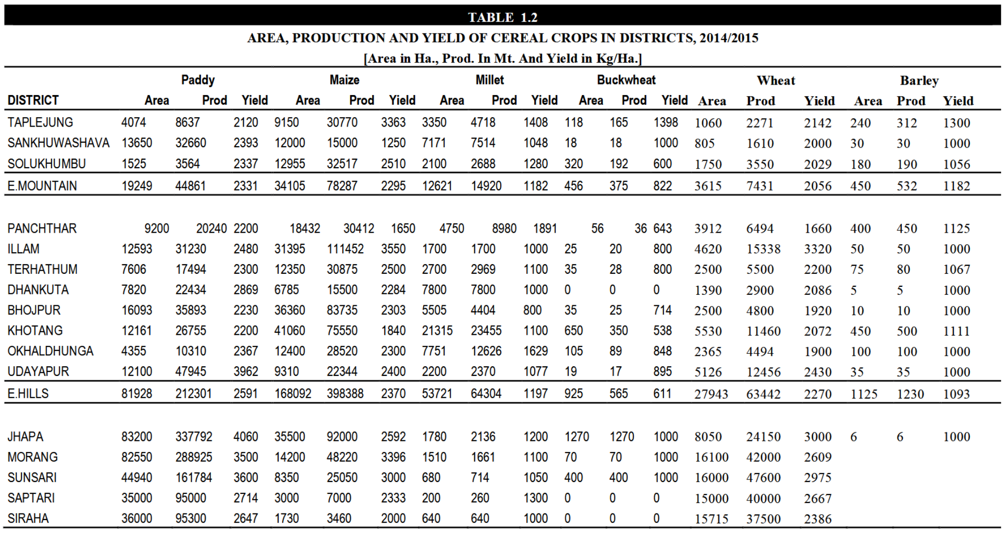


TABLE 1.2

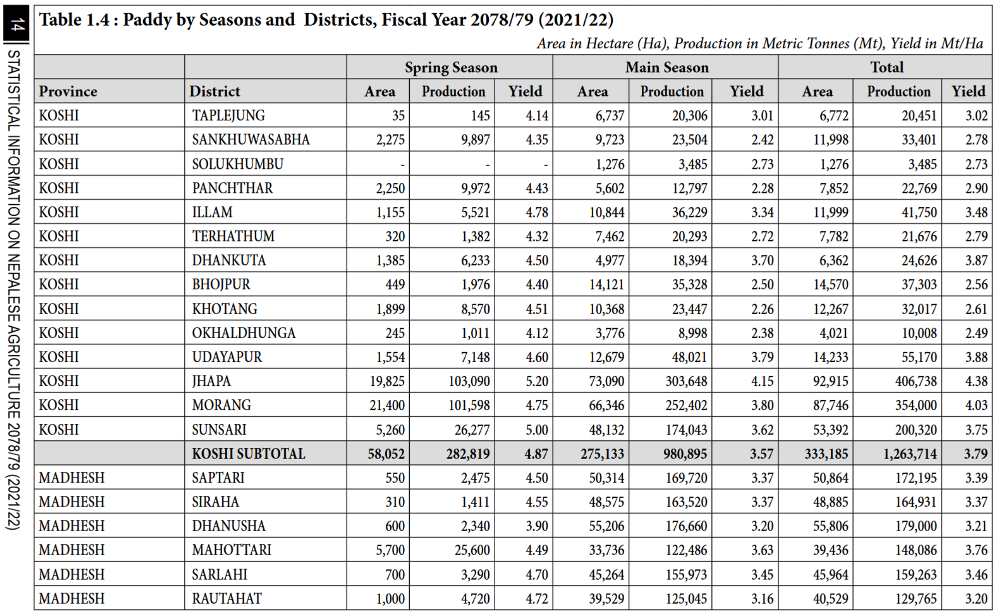
AREA, PRODUCTION AND YIELD OF CEREAL CROPS IN DISTRICTS, 2014/2015

[Area in Ha., Prod. In Mt. And Yield in Kg/Ha.]

	Paddy			Maize			Millet			Buckwheat			Wheat			Barley		
DISTRICT	Area	Prod	Yield	Area	Prod	Yield	Area	Prod	Yield	Area	Prod	Yield	Area	Prod	Yield	Area	Prod	Yield
TAPLEJUNG	4074	8637	2120	9150	30770	3363	3350	4718	1408	118	165	1398	1060	2271	2142	240	312	1300
SANKHUWASHAVA	13650	32660	2393	12000	15000	1250	7171	7514	1048	18	18	1000	805	1610	2000	30	30	1000
SOLUKHUMBU	1525	3564	2337	12955	32517	2510	2100	2688	1280	320	192	600	1750	3550	2029	180	190	1056
E.MOUNTAIN	19249	44861	2331	34105	78287	2295	12621	14920	1182	456	375	822	3615	7431	2056	450	532	1182
PANCHTHAR	9200	20240	2200	18432	30412	1650	4750	8980	1891	56	36	643	3912	6494	1660	400	450	1125
ILLAM	12593	31230	2480	31395	111452	3550	1700	1700	1000	25	20	800	4620	15338	3320	50	50	1000
TERHATHUM	7606	17494	2300	12350	30875	2500	2700	2969	1100	35	28	800	2500	5500	2200	75	80	1067
DHANKUTA	7820	22434	2869	6785	15500	2284	7800	7800	1000	0	0	0	1390	2900	2086	5	5	1000
BHOJPUR	16093	35893	2230	36360	83735	2303	5505	4404	800	35	25	714	2500	4800	1920	10	10	1000
KHOTANG	12161	26755	2200	41060	75550	1840	21315	23455	1100	650	350	538	5530	11460	2072	450	500	1111
OKHALDHUNGA	4355	10310	2367	12400	28520	2300	7751	12626	1629	105	89	848	2365	4494	1900	100	100	1000
UDAYAPUR	12100	47945	3962	9310	22344	2400	2200	2370	1077	19	17	895	5126	12456	2430	35	35	1000
E.HILLS	81928	212301	2591	168092	398388	2370	53721	64304	1197	925	565	611	27943	63442	2270	1125	1230	1093
JHAPA	83200	337792	4060	35500	92000	2592	1780	2136	1200	1270	1270	1000	8050	24150	3000	6	6	1000
MORANG	82550	288925	3500	14200	48220	3396	1510	1661	1100	70	70	1000	16100	42000	2609			
SUNSARI	44940	161784	3600	8350	25050	3000	680	714	1050	400	400	1000	16000	47600	2975			
SAPTARI	35000	95000	2714	3000	7000	2333	200	260	1300	0	0	0	15000	40000	2667			
SIRAHA	36000	95300	2647	1730	3460	2000	640	640	1000	0	0	0	15715	37500	2386			

Figure 2. *Example photo of a table with structural inconsistencies*

Note. This is the publication from 2015 that details crop yield from 2014-2015



14 STATISTICAL INFORMATION ON NEPALESE AGRICULTURE 2078/79 (2021/22)

Table 1.4 : Paddy by Seasons and Districts, Fiscal Year 2078/79 (2021/22)

Area in Hectare (Ha), Production in Metric Tonnes (Mt), Yield in Mt/Ha

		Spring Season			Main Season			Total		
Province	**District**	**Area**	**Production**	**Yield**	**Area**	**Production**	**Yield**	**Area**	**Production**	**Yield**
KOSHI	TAPLEJUNG	35	145	4.14	6,737	20,306	3.01	6,772	20,451	3.02
KOSHI	SANKHUWASABHA	2,275	9,897	4.35	9,723	23,504	2.42	11,998	33,401	2.78
KOSHI	SOLUKHUMBU	-	-	-	1,276	3,485	2.73	1,276	3,485	2.73
KOSHI	PANCHTHAR	2,250	9,972	4.43	5,602	12,797	2.28	7,852	22,769	2.90
KOSHI	ILLAM	1,155	5,521	4.78	10,844	36,229	3.34	11,999	41,750	3.48
KOSHI	TERHATHUM	320	1,382	4.32	7,462	20,293	2.72	7,782	21,676	2.79
KOSHI	DHANKUTA	1,385	6,233	4.50	4,977	18,394	3.70	6,362	24,626	3.87
KOSHI	BHOJPUR	449	1,976	4.40	14,121	35,328	2.50	14,570	37,303	2.56
KOSHI	KHOTANG	1,899	8,570	4.51	10,368	23,447	2.26	12,267	32,017	2.61
KOSHI	OKHALDHUNGA	245	1,011	4.12	3,776	8,998	2.38	4,021	10,008	2.49
KOSHI	UDAYAPUR	1,554	7,148	4.60	12,679	48,021	3.79	14,233	55,170	3.88
KOSHI	JHAPA	19,825	103,090	5.20	73,090	303,648	4.15	92,915	406,738	4.38
KOSHI	MORANG	21,400	101,598	4.75	66,346	252,402	3.80	87,746	354,000	4.03
KOSHI	SUNSARI	5,260	26,277	5.00	48,132	174,043	3.62	53,392	200,320	3.75
	KOSHI SUBTOTAL	**58,052**	**282,819**	**4.87**	**275,133**	**980,895**	**3.57**	**333,185**	**1,263,714**	**3.79**
MADHESH	SAPTARI	550	2,475	4.50	50,314	169,720	3.37	50,864	172,195	3.39
MADHESH	SIRAHA	310	1,411	4.55	48,575	163,520	3.37	48,885	164,931	3.37
MADHESH	DHANUSHA	600	2,340	3.90	55,206	176,660	3.20	55,806	179,000	3.21
MADHESH	MAHOTTARI	5,700	25,600	4.49	33,736	122,486	3.63	39,436	148,086	3.76
MADHESH	SARLAHI	700	3,290	4.70	45,264	155,973	3.45	45,964	159,263	3.46
MADHESH	RAUTAHAT	1,000	4,720	4.72	39,529	125,045	3.16	40,529	129,765	3.20

Figure 3. *Example photo of a PDF with well-structured data presentation*

Note. This is the publication from 2022 that details crop yield from 2021-2022

OCR Extraction of Text

To convert scanned PDF images into machine-readable text, the *pytesseract* library was utilized. Each page was processed using the OCR engine with specific configurations optimized for tabular data.

The configuration ensured that rows and columns are retained as much as possible during text extraction. The extracted text was then parsed into a list of rows for further processing. During this processing a few problems were encountered, they are as follows:

- Variations in table formatting required custom parsing logic., as seen in *figure 2 and figure 3*.
- OCR errors such as incorrect character recognition necessitated additional cleaning.

Data Cleaning and Normalization

The extracted text often contained artifacts like symbols, extra spaces, or misinterpreted characters. A cleaning function was implemented to normalize the data, replacing unwanted symbols and correcting common OCR errors.

For example, we realized that the most common mistake was “0” was being replaced with either “(“ or “)” brackets, we added logic to handle that. Another instance was that random symbols and extra spaces were being added in between the numerical values. Therefore, logic was added to handle those as well.

```
1. Input: A list of rows containing text data, which may
   include unwanted characters like brackets.
2. Process:
       ○  For each row in the data:
              ■  For each cell (individual text entry) in
                 the row:
                     ■  Replace opening and closing brackets
                        '(' and ')' with '0'.
                     ■  Remove any extra spaces from the
                        text.
                     ■  Keep only alphanumeric content
                        (letters and numbers), discarding any
                        special characters.
       ○  Store the cleaned row in a new list.
3. Output: A list of rows with cleaned and normalized
   text data.
```

Figure 4. *Algorithm to Clean and Normalize Extracted Text*

Alignment of Rows and Columns

The extracted data had inconsistent column lengths across rows, due to OCR errors or table irregularities, as we can see in *figure 2*, older datasets were not demarcated at all. Rows were aligned by either trimming extra columns or padding missing ones with placeholders.

```
1. Input: A list of rows with varying numbers of columns
   and an expected number of columns for each row.
2. Process:
●  For each row in the data:
       ○  If the row has fewer columns than expected, add
          extra columns with a value of '0' to match the
          expected number.
       ○  If the row has more columns than expected, remove
          the extra columns to match the expected number.
   3. Output: A new list of rows, where each row has the
      same number of columns as expected.
```

Figure 5. *Algorithm to Align Rows in a Data Table*

This approach standardized the table structure, enabling seamless integration with pandas DataFrames for further analysis.

Saving Processed Data

The final step involved saving the cleaned and aligned data into a CSV file for further processing and calculations.

Table 2. *A snapshot of the cleaned and aligned dataset*

DISTRICT	PD_P_201415	PD_A_201415
Achham	34795	16500
Arghakhanchi	25386	8189
Baglung	17715	5782
Baitadi	15680	7000
Bajhang	22769	7006
Bajura	7993	3310
Banke	102375	32500
Bara	156896	60446
Bardiya	20500	48500

Note. This table only presents the first 10 rows and data for the crop Paddy, PD_P_201415 stands for Paddy Production in the year 2014 to 2015. Similarly, PD_A_201415 stands for Paddy Area (of cultivation) in the year 2014 to 2015.

3.3 Data Processing And Feature Engineering

The collected datasets underwent extensive processing to ensure accuracy, completeness, and readiness for machine learning. This section details the preprocessing steps, including handling missing values, standardizing district changes, cleaning data, consolidating temporal information, and engineering features to enhance model performance.

3.3.1 Handling Missing and Null Values

Missing values were a significant challenge in the dataset due to gaps in historical records and reporting inconsistencies. These gaps were addressed using interpolation techniques tailored to the nature of the data.

To ensure consistency across the dataset, handling missing district data became a critical step. Over time, new districts were added, and some existing districts underwent boundary changes, causing data gaps. The methodology to address these gaps involved identifying missing districts, mapping their neighboring districts, and interpolating their data.

Identifying Missing Districts

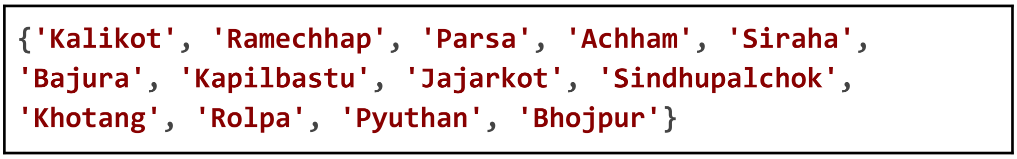
The unique district names from both crop yield and weather datasets were compared to identify disparities. The following algorithm highlights the process of finding and aligning district names:

```
1. Input: Two datasets, one containing crop yield data
   with district names and the other containing weather
   data with district names.
2. Process:
     ○ Extract the list of unique district names from
       both datasets.
     ○ Compare the two lists to identify districts that
       are present in one dataset but missing from the
       other.
          ■ Identify districts missing in the weather
            dataset (districts present in crop yield
            but not in weather).
          ■ Identify districts missing in the crop
            dataset (districts present in weather but
            not in crop yield).
3. Output: Print the list of missing districts in both
```

```
      datasets.
```

Figure 6. *Algorithm for Identifying Missing Districts*

The output provided a list of districts missing in one dataset but present in the other. To standardize naming conventions, district names were reviewed and manually updated to ensure consistency across datasets. After naming conventions were standardized, the following districts were missing from the dataset.

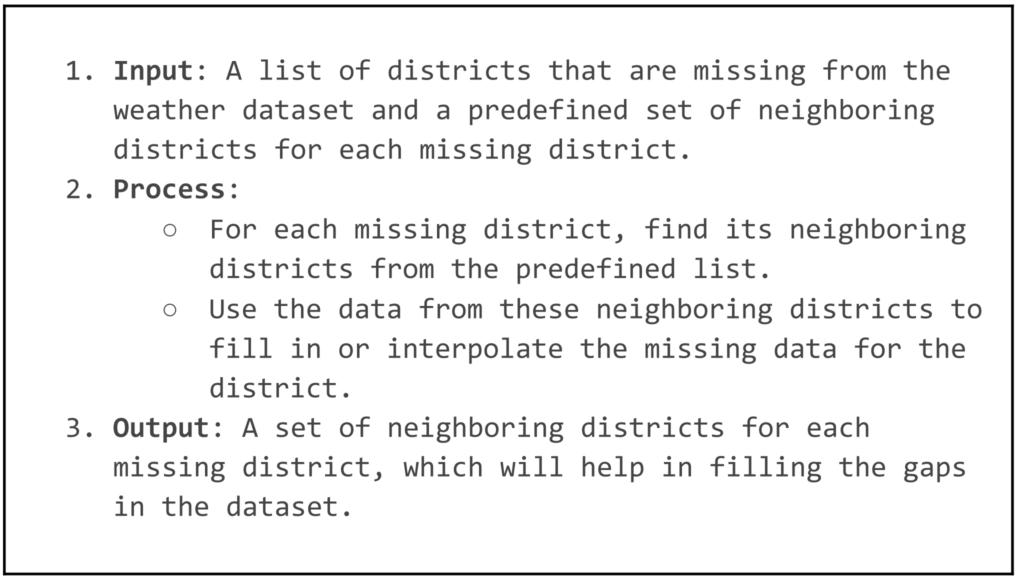


```
{'Kalikot', 'Ramechhap', 'Parsa', 'Achham', 'Siraha',
'Bajura', 'Kapilbastu', 'Jajarkot', 'Sindhupalchok',
'Khotang', 'Rolpa', 'Pyuthan', 'Bhojpur'}
```

Figure 7. *A set with the missing datasets*

Interpolating Missing District Data

For districts missing from the weather dataset, data were interpolated using neighboring districts. The neighbors of the missing districts from the previous list were determined.



```
1. Input: A list of districts that are missing from the
   weather dataset and a predefined set of neighboring
   districts for each missing district.
2. Process:
      ○ For each missing district, find its neighboring
        districts from the predefined list.
      ○ Use the data from these neighboring districts to
        fill in or interpolate the missing data for the
        district.
3. Output: A set of neighboring districts for each
   missing district, which will help in filling the gaps
   in the dataset.
```

Figure 8. *Algorithm for Interpolating Missing District Data*

Crop Yield Data

Missing data for a district and year were then filled using the above information. An algorithm explaining the logic is given.

```
1. Input: A dictionary of missing districts with their
   neighboring districts and crop yield data for each
   district and year.
2. Process:
      ○ For each missing district and its associated
        neighboring districts:
           ■ For each year, filter the crop yield data
             for the neighboring districts.
           ■ If data for the neighboring districts is
             available for the given year, calculate the
             mean of the relevant numerical columns
             (e.g., 'AREA', 'PRODUCTION', 'YIELD').
           ■ Round the mean values and use them to fill
             in the missing data for the current
             district.
           ■ Create a new row for the missing district
             with the interpolated values for that year.
3. Output: A list of new rows with the interpolated data
   for the missing districts.
```

Figure 9. *Algorithm for Filling Missing Crop Yield Data*

Weather Data

The following snippet illustrates the logic used to calculate average weather data for missing districts using the neighbors list.

```
1. Input: A dictionary of missing districts with their
   neighboring districts and weather data for each
   district and year.
2. Process:
      ○ For each missing district and its associated
```

```
            neighboring districts:
               ■  For each year, filter the weather data for
                  the neighboring districts.
               ■  If data for the neighboring districts is
                  available for the given year, calculate the
                  mean of the relevant numerical columns
                  (e.g., temperature, rainfall, etc.).
               ■  Round the mean values and use them to fill
                  in the missing data for the current
                  district.
               ■  Create a new row for the missing district
                  with the interpolated values for that year.
3. **Output**: A list of new rows with the interpolated
   weather data for the missing districts.
```

Figure 10. *Algorithm for Filling Missing Weather Data*

This approach ensured that missing districts were populated with realistic data derived from neighboring regions. A final check verified that all districts were accounted for, and results were saved into a consolidated file for further analysis.

3.3.2 Cleaning Data

Errors introduced during data extraction were addressed to ensure consistency.

Outlier Detection and Removal

Extreme values were flagged using the interquartile range (IQR) method.

```
1. Input: A dataset with a column 'Value' containing
   numerical data.
2. Process:
      ○  Calculate the first quartile (Q1) and third
         quartile (Q3) of the 'Value' column.
      ○  Compute the interquartile range (IQR) as the
         difference between Q3 and Q1.
      ○  Identify outliers as values that fall outside the
         range of:
            ■ 'Q1 - 1.5 * IQR' (lower bound)
            ■ 'Q3 + 1.5 * IQR' (upper bound)
        ○ Remove any rows where the 'Value' falls outside
          this range.
3. Output: A dataset with outliers removed, keeping only
   the rows where 'Value' is within the acceptable range.
```

Figure 11. *Algorithm for Detecting and Removing Outliers Using IQR*

3.3.3 Consolidating Temporal Data into Seasonal Features

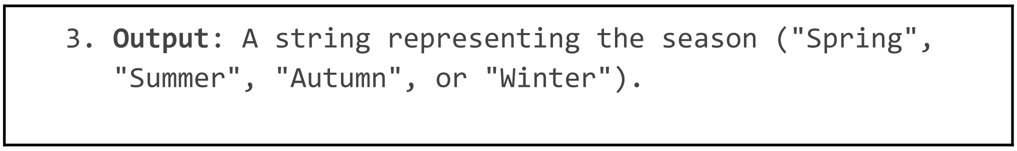
To obtain seasonal weather factors from the monthly weather data, we implemented a systematic approach that involved classifying each month into its respective season and then aggregating the data accordingly. The process was carried out in the following steps.

Classifying Months into Seasons

The first step involved categorizing the months of the year into one of the four seasons: Winter, Spring, Summer, and Autumn. This classification was based on standard meteorological definitions as follows:

- Spring includes the months of March, April, and May.
- Summer includes June, July, August, and September.
- Autumn encompasses October and November.
- Winter covers December, January, and February.

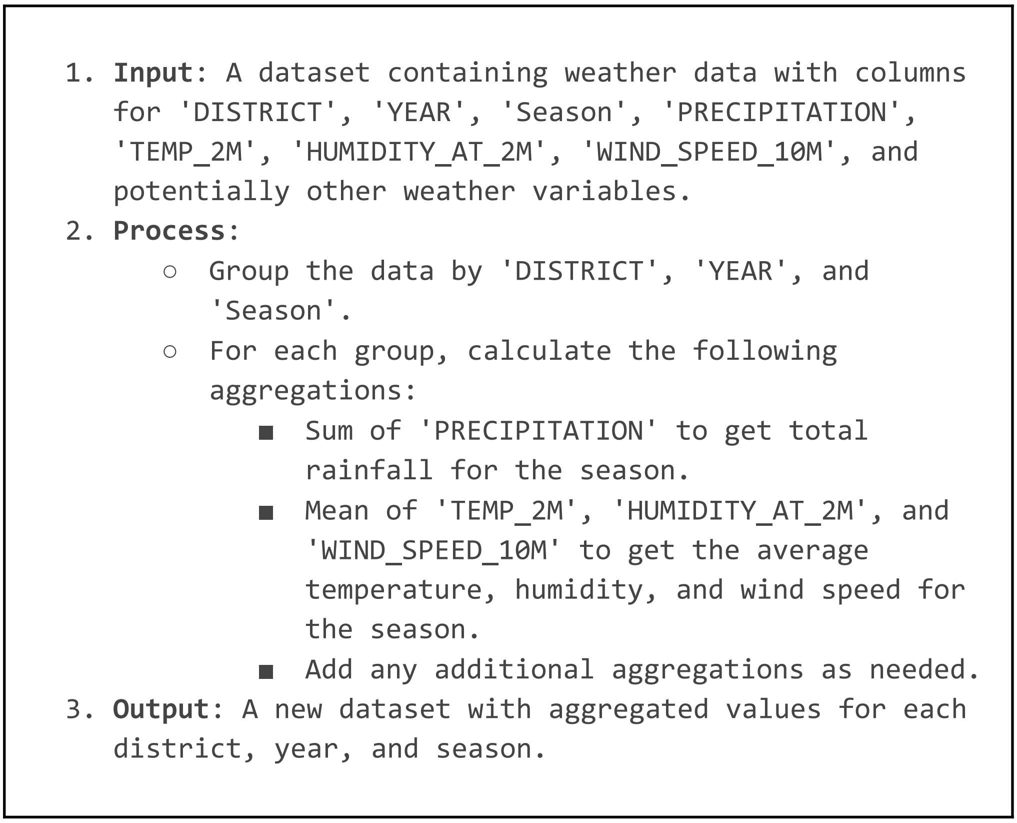
```
1. Input: A numerical value representing the month (1 for
   January, 2 for February, etc.).
2. Process:
        ○ If the month is 3, 4, or 5, classify it as
          "Spring".
        ○ If the month is 6, 7, 8, or 9, classify it as
          "Summer".
        ○ If the month is 10 or 11, classify it as
          "Autumn".
        ○ Otherwise, classify it as "Winter" (for December,
          January, and February).
```



```
3. Output: A string representing the season ("Spring",
   "Summer", "Autumn", or "Winter").
```

Figure 12. *Algorithm for Classifying Seasons Based on Month*

Then the average values of each parameter in the weather dataset was calculated across each respective season.



```
1. Input: A dataset containing weather data with columns
   for 'DISTRICT', 'YEAR', 'Season', 'PRECIPITATION',
   'TEMP_2M', 'HUMIDITY_AT_2M', 'WIND_SPEED_10M', and
   potentially other weather variables.
2. Process:
      ○ Group the data by 'DISTRICT', 'YEAR', and
        'Season'.
      ○ For each group, calculate the following
        aggregations:
            ■ Sum of 'PRECIPITATION' to get total
              rainfall for the season.
            ■ Mean of 'TEMP_2M', 'HUMIDITY_AT_2M', and
              'WIND_SPEED_10M' to get the average
              temperature, humidity, and wind speed for
              the season.
            ■ Add any additional aggregations as needed.
3. Output: A new dataset with aggregated values for each
   district, year, and season.
```

Figure 13. *Algorithm for Aggregating Seasonal Weather Data*

This process left us with seasonal weather inferences such as total precipitation, average temperature, average humidity, etc. but for each of the four seasons that we experience in Nepal (Ghimire, 2019).

Table 3. *Table with aggregated seasonal weather features*

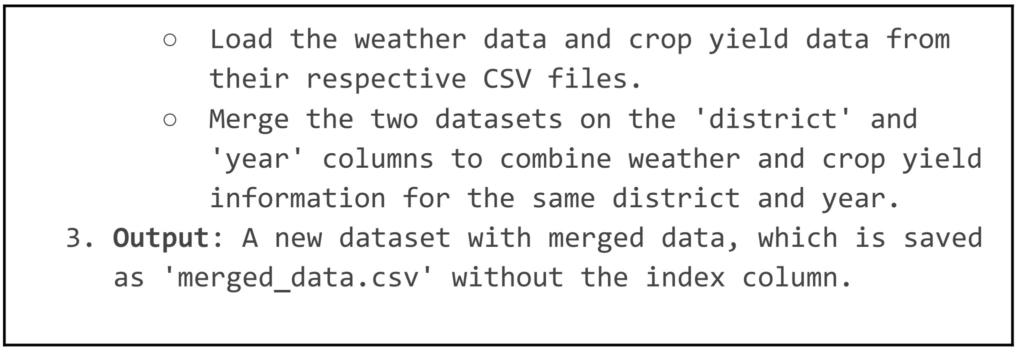
District	Year	Autumn_humidity	Spring_humidity	Summer_humidity	Winter_humidity
Arghakhanchi	1981	9.14	7.46	17.06	5.08
Arghakhanchi	1982	8.49	7.1	16.04	4.76
Arghakhanchi	1983	9.56	5.68	15.39	3.7
Arghakhanchi	1984	7.93	5.29	17	3.82
Arghakhanchi	1985	10.56	4.74	15.76	4.41
Arghakhanchi	1986	6.98	5.07	15.3	4.42
Arghakhanchi	1987	8.76	5.15	15.41	4.05
Arghakhanchi	1988	8.86	5.58	17.44	4.36
Arghakhanchi	1989	7.04	4.82	16.89	3.64

Note. This snapshot only contains the first 10 rows and humidity feature for the 4 seasons. The entire dataset contains data for each district across 1981 to 2022 with various other features like temperature and rainfall.

3.4 Dataset Consolidation

To continue with further analysis, the two distinct datasets were combined. We used the columns DISTRICT and YEAR to merge the weather and crop yield dataset, this kept the rows that had corresponding values in both datasets and dropped any ambiguities. This allowed us to feed the data directly into the model later. This process was completed using the following logic.

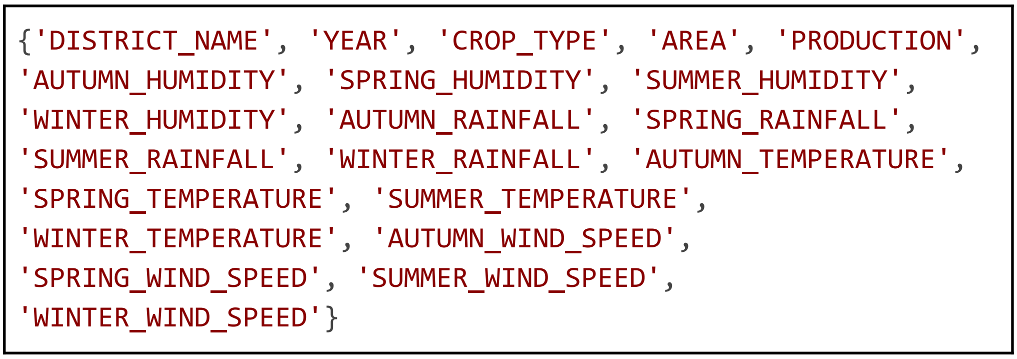
```
1. Input: Two datasets, one containing weather data
   ('weather_data.csv') and the other containing crop
   yield data ('crop_yield_data.csv'). Both datasets
   should have columns for 'district' and 'year'.
2. Process:
```



```
          ○  Load the weather data and crop yield data from
             their respective CSV files.
          ○  Merge the two datasets on the 'district' and
             'year' columns to combine weather and crop yield
             information for the same district and year.
    3. Output: A new dataset with merged data, which is saved
       as 'merged_data.csv' without the index column.
```

Figure 14. *Algorithm for Merging Weather and Crop Yield Datasets*

At this point, the dataset contained the following columns overall.



```
{'DISTRICT_NAME', 'YEAR', 'CROP_TYPE', 'AREA', 'PRODUCTION',
'AUTUMN_HUMIDITY', 'SPRING_HUMIDITY', 'SUMMER_HUMIDITY',
'WINTER_HUMIDITY', 'AUTUMN_RAINFALL', 'SPRING_RAINFALL',
'SUMMER_RAINFALL', 'WINTER_RAINFALL', 'AUTUMN_TEMPERATURE',
'SPRING_TEMPERATURE', 'SUMMER_TEMPERATURE',
'WINTER_TEMPERATURE', 'AUTUMN_WIND_SPEED',
'SPRING_WIND_SPEED', 'SUMMER_WIND_SPEED',
'WINTER_WIND_SPEED'}
```

Figure 15. *A list with all the columns of the dataset*

3.5 Model Implementation

This section covers the process of model implementation, consisting of data pre-processing, model training and tuning, model evaluation and visualization.

3.5.1 Data Pre-Processing

The preprocessing steps included before the data was fed into the model were as follows:

One hot encoding

The categorical DISTRICT column was transformed using one-hot encoding to account for its influence on yield. Converting the string to an integer value allows us to include the district value into the model. This will allow us to identify any significant relationship between a certain district and a crop's yield.

Feature selection

Features (independent variables) were separated from the target variable (crop yield). Independent variables included the district and all the weather information. The year column was dropped entirely, as we are not concerned with the time series analysis of the dataset.

Train-test split

The data was divided into training and testing sets using a split ratio of 80% for training and 20% for testing. This split ensures the models are trained on a representative portion of the data and evaluated on unseen data.

Scaling for SVR

Standard scaling was applied to the training data for the Support Vector Regression (SVR) model to improve its performance, as the model performance degrades significantly if we do not scale the training data.

The algorithm demonstrating the above is as follows:

```
1. Input: A dataset with columns 'DISTRICT_NAME', 'YEAR',
   'CROP_TYPE', 'PRODUCTION', and other features.
2. Process:
      ○ Apply one-hot encoding to the 'DISTRICT_NAME'
        column to convert categorical district names into
        binary columns (excluding the first column to
        avoid multicollinearity).
      ○ Drop the columns 'YEAR', 'CROP_TYPE', and
        'PRODUCTION', as they are not required for the
        feature set, and separate the remaining columns
        into features (X) and target (y).
      ○ Split the data into training (80%) and testing
        (20%) sets using a random state for
        reproducibility.
      ○ Scale the features using a standard scaler to
        ensure that all features have a similar range,
          which is especially important for models like
          SVR.
   3. Output: Scaled training and testing feature sets
      (X_train_scaled and X_test_scaled), along with the
      corresponding target variables (y_train and y_test).
```

Figure 16. *Algorithm for Preprocessing Data for Model Training*

3.5.2 Model Training And Tuning

Rationale Behind Model Selection

Three distinct machine learning models were selected for this study, each chosen for its unique strengths and suitability for different aspects of the crop yield prediction task.

Linear Regression: This model serves as a foundational baseline. It provides insights into the linear relationships between weather variables and crop yield, offering a simple and interpretable model for initial analysis (Dongarra et al., 2011).

Random Forest: As an ensemble learning method, Random Forest is well-suited for handling complex, non-linear relationships often present in agricultural systems. Its ability to capture intricate interactions between various factors influencing crop yield makes it a strong contender for accurate predictions (Jeong et al., 2016; Schonlau & Zou, 2020).

Support Vector Regression (SVR): A version of Support Vector Machines (SVM), SVR is known for its robustness to outliers and its effectiveness in handling non-linear relationships. This makes it a suitable candidate for modeling the potential complexities and variations observed in real-world agricultural data (Mountrakis et al., 2011; Priyadharshini et al., 2022).

This combination of models allows for a comprehensive exploration of different modeling approaches, providing valuable insights into the factors driving crop yield variability and identifying the most effective model for accurate predictions in the Nepalese context. These were also amongst the most popular models used for crop yield prediction (Van Klompenburg et al., 2020).

Hyperparameter tuning

Grid Search with *Cross-Validation* was employed to determine the optimal hyperparameter values for the Random Forest and SVR model. The parameters explored included n_estimators (number of trees) and max_depth (maximum depth of each tree) for Random Forest and, C (penalty parameter) and epsilon (tolerance for errors) for SVR.

The Random Forest and SVR models were then trained using the values that we obtained from the grid search process.

```
1. Input:
      ○  For Random Forest: A parameter grid with values
         for 'n_estimators' (number of trees) and
         'max_depth' (maximum depth of each tree).
      ○  For SVR: A parameter grid with values for 'C'
         (regularization parameter) and 'epsilon' (margin
         of tolerance for error).
2. Process:
      ○  Define the parameter grids for both models
         (Random Forest and SVR).
      ○  Create GridSearchCV objects for each model:
            ■  For Random Forest, use the
               RandomForestRegressor estimator with a
               cross-validation (cv=5) and scoring based
               on the negative mean squared error.
            ■  For SVR, use the SVR estimator with a
               linear kernel, cross-validation (cv=5), and
               scoring based on the negative mean squared
               error.
3. Output:
      ○  grid_search_rf and grid_search_svr objects that
         perform hyperparameter tuning to find the best
         set of parameters for each model based on the
         given grids.
```

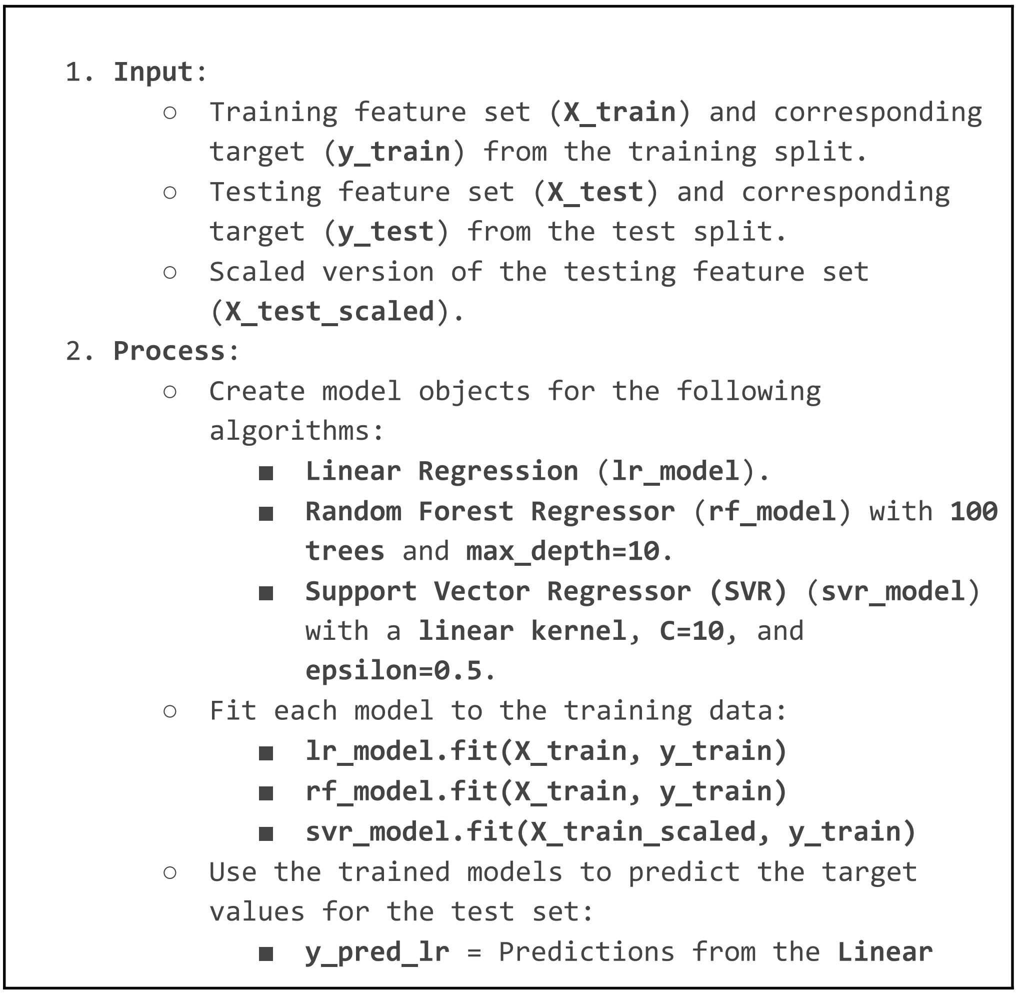
Figure 17. *Algorithm for Hyperparameter Tuning Using GridSearchCV*

After running the script, the following values were obtained to be used as the values for hyperparameters.

Table 4. *Values for hyperparameters obtained from Grid Search*

Model	Parameter	Best Value
Random Forest	n_estimators	100
Random Forest	max_depth	10
SVR	kernel	linear
SVR	C	10
SVR	epsilon	0.5

After completing all of the above, we moved on to training the actual model. An algorithm explaining it is given below.



```
1. Input:
      ○ Training feature set (X_train) and corresponding
        target (y_train) from the training split.
      ○ Testing feature set (X_test) and corresponding
        target (y_test) from the test split.
      ○ Scaled version of the testing feature set
        (X_test_scaled).
2. Process:
      ○ Create model objects for the following
        algorithms:
            ■ Linear Regression (lr_model).
            ■ Random Forest Regressor (rf_model) with 100
              trees and max_depth=10.
            ■ Support Vector Regressor (SVR) (svr_model)
              with a linear kernel, C=10, and
              epsilon=0.5.
      ○ Fit each model to the training data:
            ■ lr_model.fit(X_train, y_train)
            ■ rf_model.fit(X_train, y_train)
            ■ svr_model.fit(X_train_scaled, y_train)
      ○ Use the trained models to predict the target
        values for the test set:
            ■ y_pred_lr = Predictions from the Linear
```

```
                 Regression model.
              ■  y_pred_rf = Predictions from the Random
                 Forest Regressor model.
              ■  y_pred_svr = Predictions from the SVR
                 model.
3. Output:
      ○  y_pred_lr, y_pred_rf, and y_pred_svr, which are
         the predicted values for the test set by each
         respective model.
```

Figure 18. *Algorithm for Training Models and Making Predictions*

The predictions, or y_pred_lr, y_pred_rf and y_pred_svr stand for the array that contains the values which the model has predicted, after being trained, on the testing data that we separated at the beginning. This will be used to calculate the model performance metric later on.

3.5.3 Model Evaluation Metrics

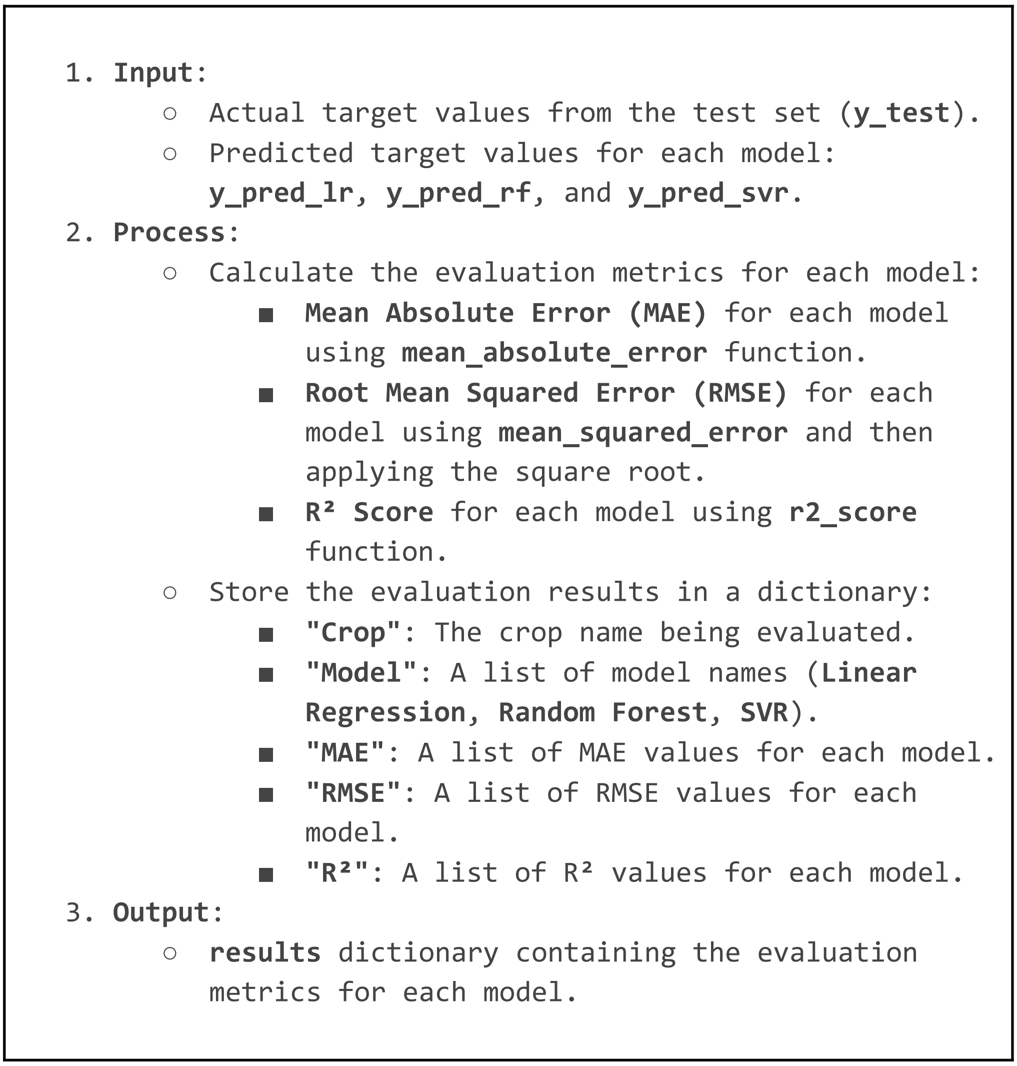
As all the models were used for a regression based prediction task, the performance of the trained models was evaluated using regression-based metrics.

Mean Absolute Error (MAE): MAE measures the average absolute difference between the actual and predicted crop yields. Lower MAE values indicate higher accuracy, as it minimizes the average prediction error. It is also considered to be ones of the fundamental measures of model performance, often considered to be more natural and better than RMSE (Willmott & Matsuura, 2005).

Root Mean Squared Error (RMSE): RMSE provides a measure of the standard deviation of the residuals. It gives more weight to larger errors, making it sensitive to outliers. Lower RMSE values signify better model performance in terms of minimizing prediction errors. It is considered an inferior measure to MAE for model performance evaluation by Willmott &

Matsuura (2005), but according to Hodson (2022), depending on the nature of the dataset, RMSE can provide insights that are missed by MAE.

R-squared (R²): R-squared represents the proportion of variance in the actual crop yield that is explained by the model's predictions. The value of R-squared ranges from 0 to 1 or 0% to 100%. A higher R-squared value indicates a better fit, suggesting that the model effectively captures the underlying patterns in the data. It is considered as being more informative than the above two and best for analysis of regression models (Chicco et al., 2021).
An algorithm that demonstrates the usage of these metrics is provided.



```
1. Input:
      ○ Actual target values from the test set (y_test).
      ○ Predicted target values for each model:
        y_pred_lr, y_pred_rf, and y_pred_svr.
2. Process:
      ○ Calculate the evaluation metrics for each model:
            ■ Mean Absolute Error (MAE) for each model
              using mean_absolute_error function.
            ■ Root Mean Squared Error (RMSE) for each
              model using mean_squared_error and then
              applying the square root.
            ■ R² Score for each model using r2_score
              function.
      ○ Store the evaluation results in a dictionary:
            ■ "Crop": The crop name being evaluated.
            ■ "Model": A list of model names (Linear
              Regression, Random Forest, SVR).
            ■ "MAE": A list of MAE values for each model.
            ■ "RMSE": A list of RMSE values for each
              model.
            ■ "R²": A list of R² values for each model.
3. Output:
      ○ results dictionary containing the evaluation
        metrics for each model.
```

Figure 19. *Algorithm for Model Evaluation*

As discussed earlier, for each measure MAE, RMSE and R-squared, the predicted values (y_pred_model) given by the trained models on the split testing data, and actual yield values (y_test), were utilized to measure the model performance.

3.5.4 Result Visualization

To gain deeper insights into model performance and identify key influencing factors, several visualizations were employed:

Actual vs. Predicted Yield Scatter Plot: This plot visually compared the actual crop yields with the predicted yields for each model. Deviations from the ideal 45-degree line (representing perfect predictions) provided a visual assessment of model accuracy and potential biases.

The algorithm responsible for this plot is given along.

```
1. Input:
      ○ y_test: Actual crop yield values from the test
        set.
      ○ y_pred_lr: Predicted crop yield values using
        Linear Regression.
      ○ y_pred_rf: Predicted crop yield values using
        Random Forest.
      ○ y_pred_svr: Predicted crop yield values using
        Support Vector Regression (SVR).
2. Process:
      ○ Initialize a plot with a size of 10x6 inches.
      ○ Create scatter plots for each model:
            ■ Plot actual vs predicted values for Linear
              Regression with a label and alpha
              transparency of 0.6.
            ■ Plot actual vs predicted values for Random
              Forest with a label and alpha transparency
              of 0.6.
            ■ Plot actual vs predicted values for SVR
              with a label, alpha transparency of 0.6,
              and marker "x".
      ○ Add a red dashed line representing the Perfect
```

```
            Fit (ideal 45-degree line) from the minimum to
            maximum of y_test.
          ○ Set the title of the plot using the crop_name.
          ○ Label the x-axis as "Actual Yield" and the y-axis
            as "Predicted Yield".
          ○ Include a legend to label the three models and
            the perfect fit line.
          ○ Display the plot.
   3. Output:
          ○ Scatter plot comparing the actual vs predicted
            crop yields for each model. The Perfect Fit line
            allows for visual assessment of the prediction
            accuracy and potential biases for each model.


plt.show()
```

Figure 20. *Algorithm for Actual vs Predicted Yield Scatter Plot*

Residual Plots: Residual plots were generated to analyze the distribution of errors for each model. Patterns in the residuals, such as non-randomness or heteroscedasticity (unequal variance of errors within the observations), can indicate potential model limitations or areas for improvement.

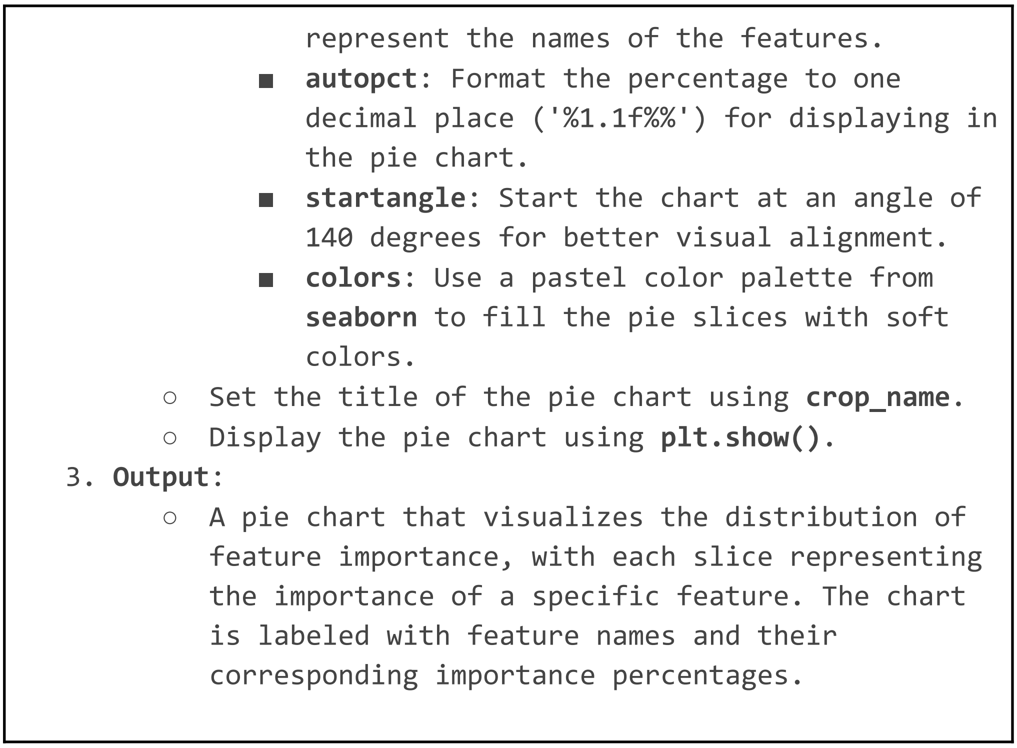
```
   1. Input:
          ○ y_test: Actual crop yield values from the test
            set.
          ○ y_pred_rf: Predicted crop yield values using
            Random Forest.
   2. Process:
          ○ Compute the residuals (errors) for the Random
            Forest model by subtracting y_pred_rf from
            y_test.
          ○ Initialize a plot with a size of 10x6 inches.
          ○ Create a scatter plot with:
                ■ x-axis: Predicted values (y_pred_rf).
                ■ y-axis: Residuals (y_test - y_pred_rf).
                ■ Label the plot as Random Forest with alpha
                  transparency of 0.6.
          ○ Add a horizontal red dashed line at y=0 to
            represent the Zero Error Line, which is the
```

```
         baseline indicating no error (perfect
         prediction).
      ○  Set the title of the plot using crop_name.
      ○  Label the x-axis as "Predicted Yield" and the
         y-axis as "Residuals".
      ○  Include a legend to label the Random Forest and
         the Zero Error Line.
      ○  Display the plot.
3. Output:
      ○  A residual plot showing the predicted yields on
         the x-axis and residuals (errors) on the y-axis.
         The Zero Error Line helps evaluate the model's
         prediction performance, with points dispersed
         around this line indicating prediction accuracy
         or bias.
```

Figure 21. *Algorithm for Residual Plot (Random Forest)*

Feature Importance (Random Forest): A pie chart was used to visualize the relative importance of different features as determined by the Random Forest model. This analysis provides valuable insights into the key drivers of crop yield variability, guiding future research and targeted interventions (Schonlau & Zou, 2020).

```
1. Input:
      ○  filtered_importance_df: A DataFrame containing
         the feature names and their corresponding
         normalized importance percentages.
      ○  crop_name: Name of the crop, used for the title
         of the plot.
2. Process:
      ○  Initialize a plot with a size of 8x8 inches.
      ○  Create a pie chart using matplotlib.pyplot.pie
         with the following parameters:
            ■  data: Values from
               filtered_importance_df['Normalized
               Importance (%)'], which represent the
               importance percentages of each feature.
            ■  labels: Labels for the pie sections from
               filtered_importance_df['Feature'], which
```



```
                  represent the names of the features.
               ■  autopct: Format the percentage to one
                  decimal place ('%1.1f%%') for displaying in
                  the pie chart.
               ■  startangle: Start the chart at an angle of
                  140 degrees for better visual alignment.
               ■  colors: Use a pastel color palette from
                  seaborn to fill the pie slices with soft
                  colors.
         ○  Set the title of the pie chart using crop_name.
         ○  Display the pie chart using plt.show().
   3. Output:
         ○  A pie chart that visualizes the distribution of
            feature importance, with each slice representing
            the importance of a specific feature. The chart
            is labeled with feature names and their
            corresponding importance percentages.
```

Figure 22. *Algorithm for Feature Importance Pie Chart (Random Forest)*

These visualizations served as valuable tools for interpreting model performance, identifying potential areas for improvement, and gaining a deeper understanding of the factors influencing crop yields in the Nepalese context.

CHAPTER 4

RESULTS

4.1 Overview Of Model Performance

This study aimed to predict the yield per area for five major cereal crops in Nepal using machine learning models. Three models—**Linear Regression, Random Forest,** and **Support Vector Regression (SVR)**—were evaluated based on their predictive performance. Metrics such as **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE),** and **R²** were used to assess their accuracy and reliability.

4.2 Model Performance Summary For Each Crop

Random Forest emerged as the best-performing model across all crops, demonstrating its ability to capture complex relationships between weather variables and crop yield. Throughout the outcomes in the above tables, RF(Random Forest) performed excellently with the minimum amount of MAE and RMSE. A comparative summary of model performances for each crop is presented below:

Table 5. *Model Performance Summary for Barley*

Model	MAE	RMSE	R^2
LR	109.68	265.9	0.87
RF	94.9	271.23	0.86
SVR	**90.07**	**247.58**	**0.88**

Table 6. *Model Performance Summary for Maize*

Model	MAE	RMSE	R^2
LR	4702.67	7608.24	0.86
RF	**3225.8**	**5948.51**	**0.92**
SVR	6087.73	11197.97	0.7

Table 7. *Model Performance Summary for Millet*

Model	MAE	RMSE	R²
LR	527.98	917.32	0.96
RF	**347.37**	**730.92**	**0.98**
SVR	482.45	1032.05	0.95

Table 8. *Model Performance Summary for Paddy*

Model	MAE	RMSE	R²
LR	11660.68	22209	0.88
RF	**8223.58**	**18021.12**	**0.92**
SVR	18006.44	37956.24	0.65

Table 9. *Model Performance Summary for Wheat*

Model	MAE	RMSE	R²
LR	4512	7427.37	0.84
RF	**2904.75**	**5675.4**	**0.91**
SVR	4848.38	9898.16	0.71

4.3 Actual Vs Predicted Yields

The scatter plots comparing the actual and predicted yields of each model for each crop is given below. The RF model's predictions closely aligned with the actual values, whereas LM and SVR struggled significantly.

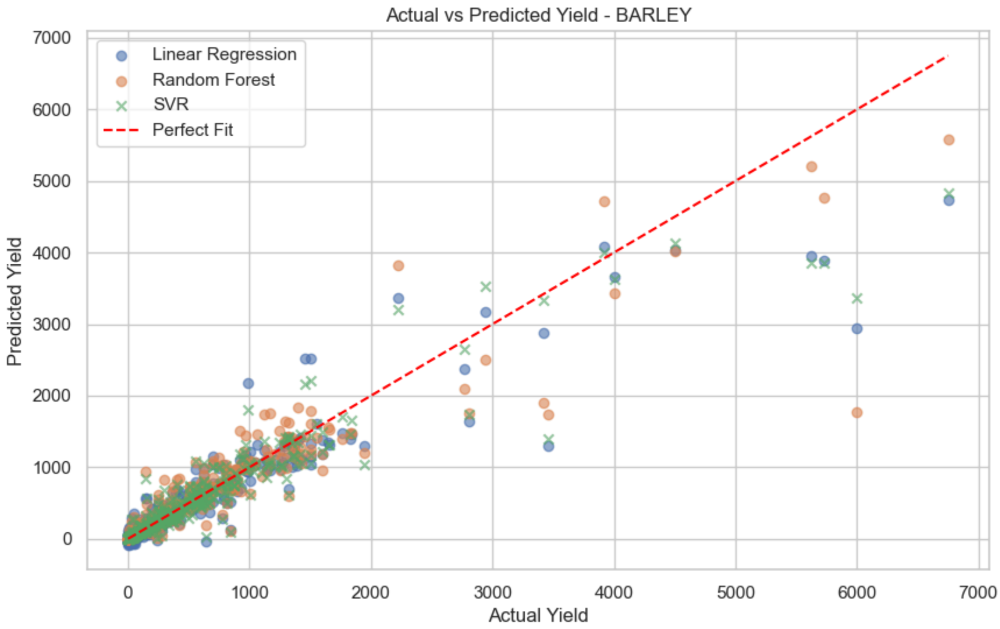


Figure 23. *Scatter plot showing Predicted vs Actual Yield for Barley*

Note. This scatter plot shows the predicted vs actual yield for the crop Barley. The X-axis represents the actual yield, and the y-axis represents the predicted yield values by the models. The central dotted line represents the perfect prediction plot.

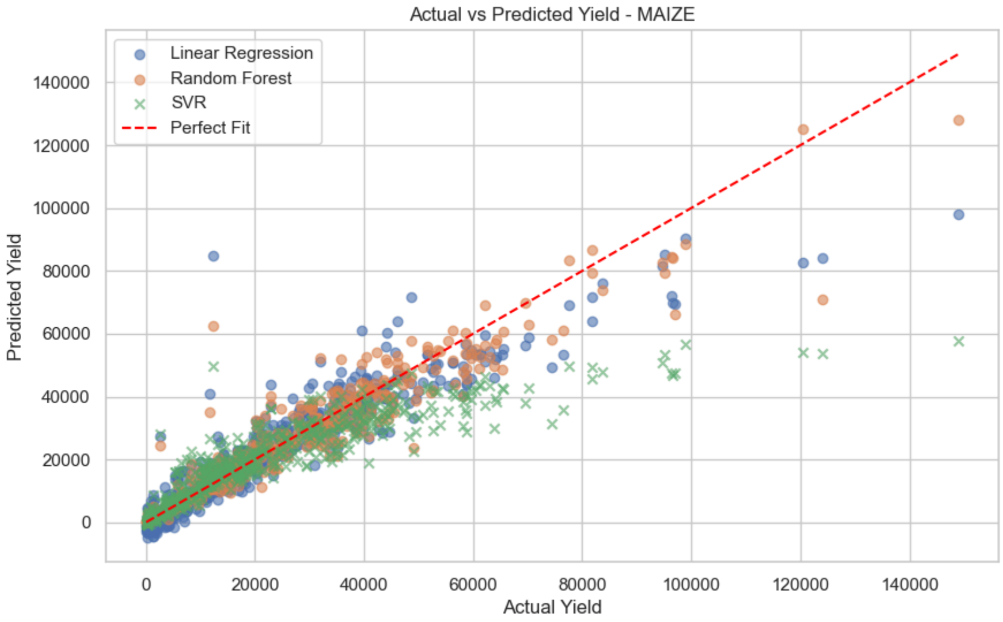


Figure 24. *Scatter plot showing Predicted vs Actual Yield for Maize*

Note. This scatter plot shows the predicted vs actual yield for the crop Maize. The X-axis represents the actual yield, and the y-axis represents the predicted yield values by the models. The central dotted line represents the perfect prediction plot.

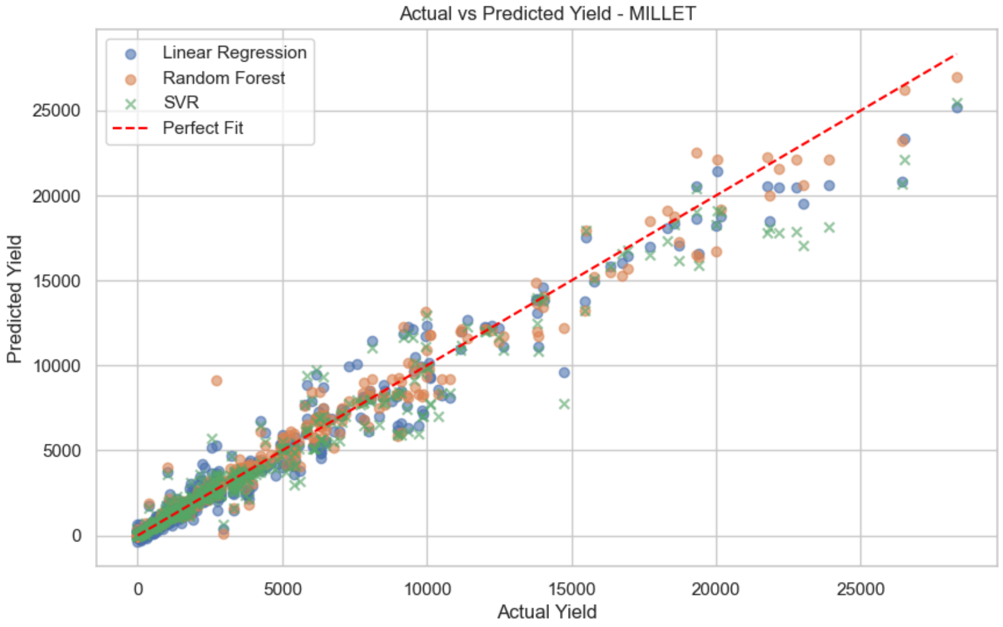


Figure 25. *Scatter plot showing Predicted vs Actual Yield for Millet*

Note. This scatter plot shows the predicted vs actual yield for the crop Millet. The X-axis represents the actual yield, and the y-axis represents the predicted yield values by the models. The central dotted line represents the perfect prediction plot.

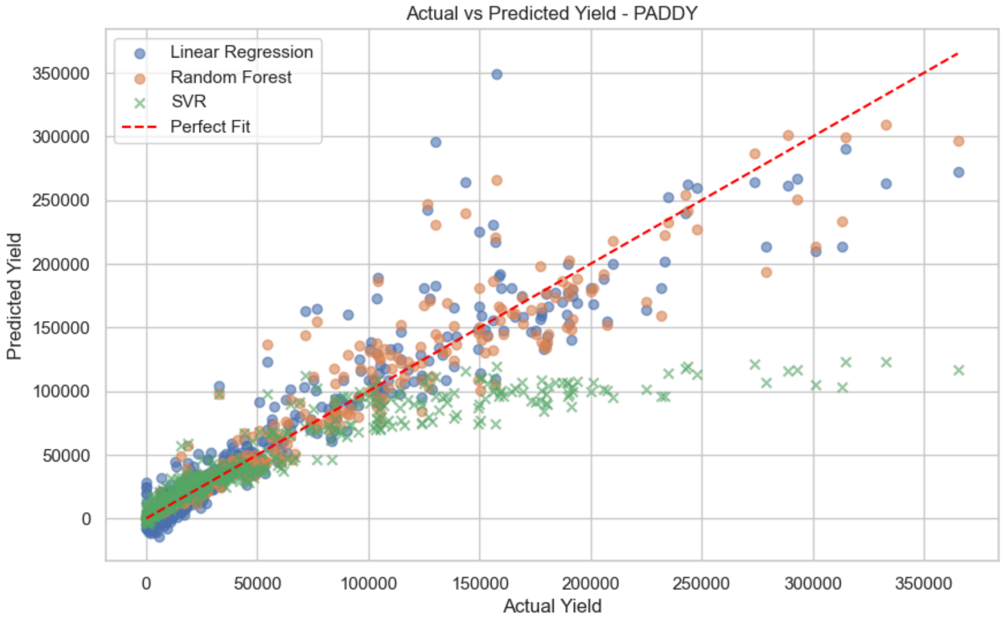


Figure 26. *Scatter plot showing Predicted vs Actual Yield for Paddy*

Note. This scatter plot shows the predicted vs actual yield for the crop Paddy. The X-axis represents the actual yield, and the y-axis represents the predicted yield values by the models. The central dotted line represents the perfect prediction plot.

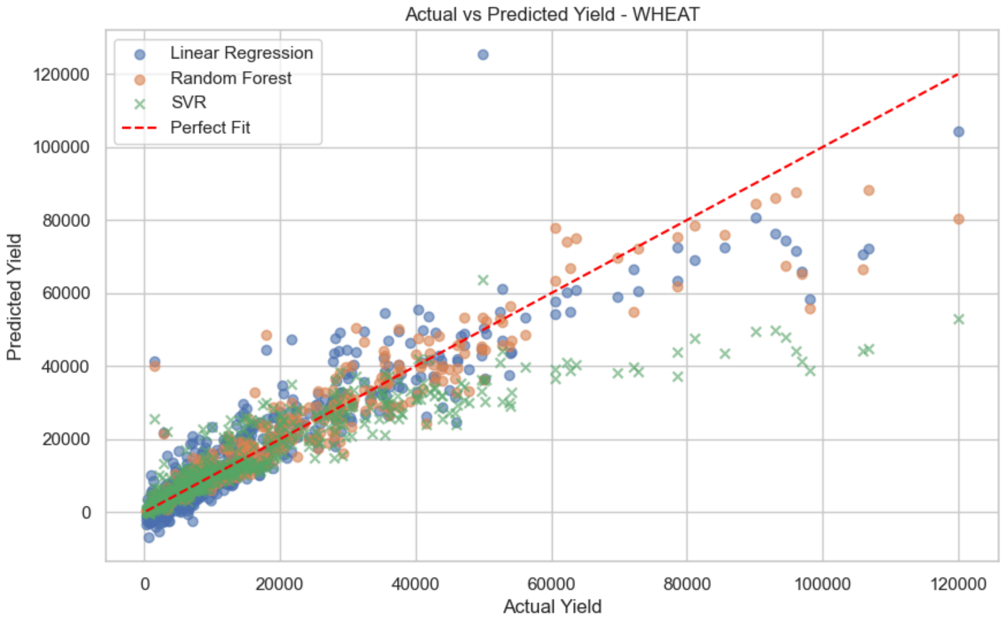


Figure 27. *Scatter plot showing Predicted vs Actual Yield for Wheat*

Note. This scatter plot shows the predicted vs actual yield for the crop Wheat. The X-axis represents the actual yield, and the y-axis represents the predicted yield values by the models. The central dotted line represents the perfect prediction plot.

4.4 Residual Plots

Residual plots were calculated alongside the scatter plots above to determine if the dataset was suitable for regression based prediction or not. Further, it allowed us to visualize any significant patterns of error or presence of **heteroscedasticity**.

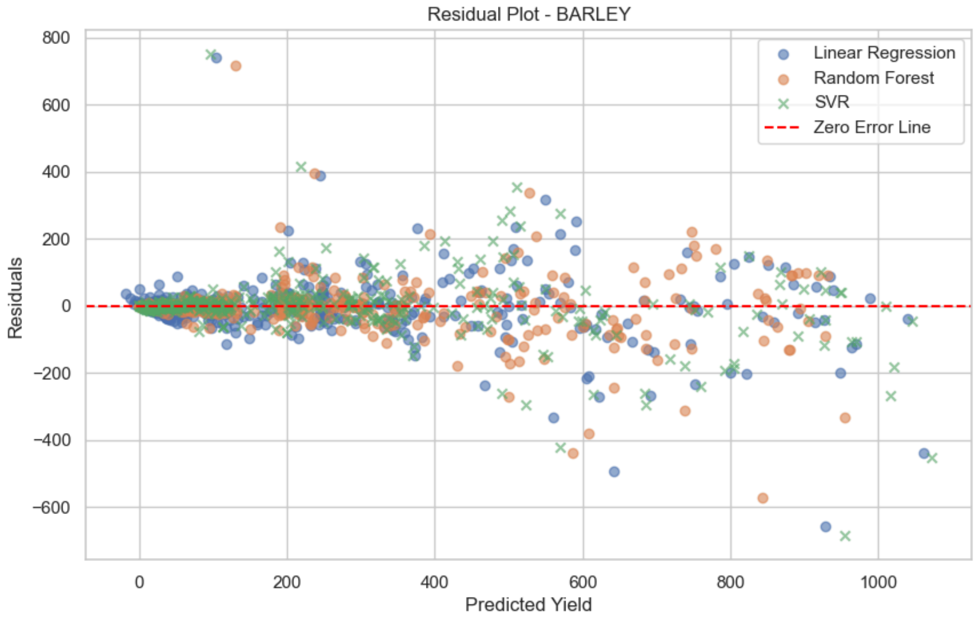


Figure 28. *Residual Plots for Barley*

Note. This residual plot shows the differences between the actual and predicted crop yield for **Paddy** using the Random Forest model. The **X-axis** represents the predicted yield, while the **Y-axis** shows the residuals. The **red dashed line** indicates the perfect prediction line (residual = 0). Points above the line indicate over-predictions, and points below indicate under-predictions. Ideally, the points should be randomly scattered around the zero line, suggesting a good model fit.

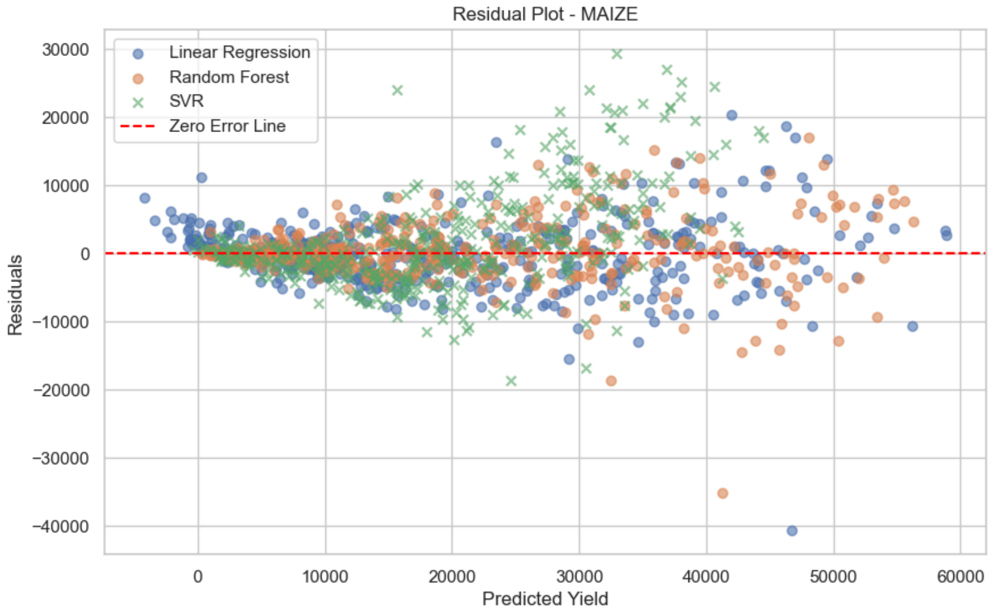


Figure 29. *Residual Plots for Maize*

Note. This residual plot shows the differences between the actual and predicted crop yield for **Maize** using the Random Forest model. The **X-axis** represents the predicted yield, while the **Y-axis** shows the residuals. The **red dashed line** indicates the perfect prediction line (residual = 0). Points above the line indicate over-predictions, and points below indicate under-predictions. Ideally, the points should be randomly scattered around the zero line, suggesting a good model fit.

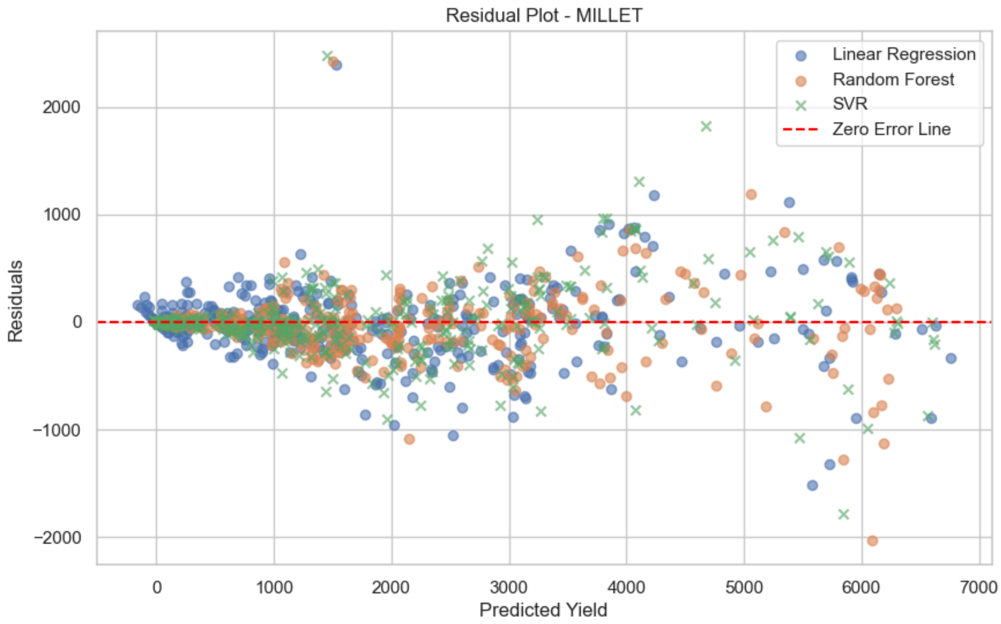


Figure 30. *Residual Plots for Millet*

Note. This residual plot shows the differences between the actual and predicted crop yield for **Millet** using the Random Forest model. The **X-axis** represents the predicted yield, while the **Y-axis** shows the residuals. The **red dashed line** indicates the perfect prediction line (residual = 0). Points above the line indicate over-predictions, and points below indicate under-predictions. Ideally, the points should be randomly scattered around the zero line, suggesting a good model fit.

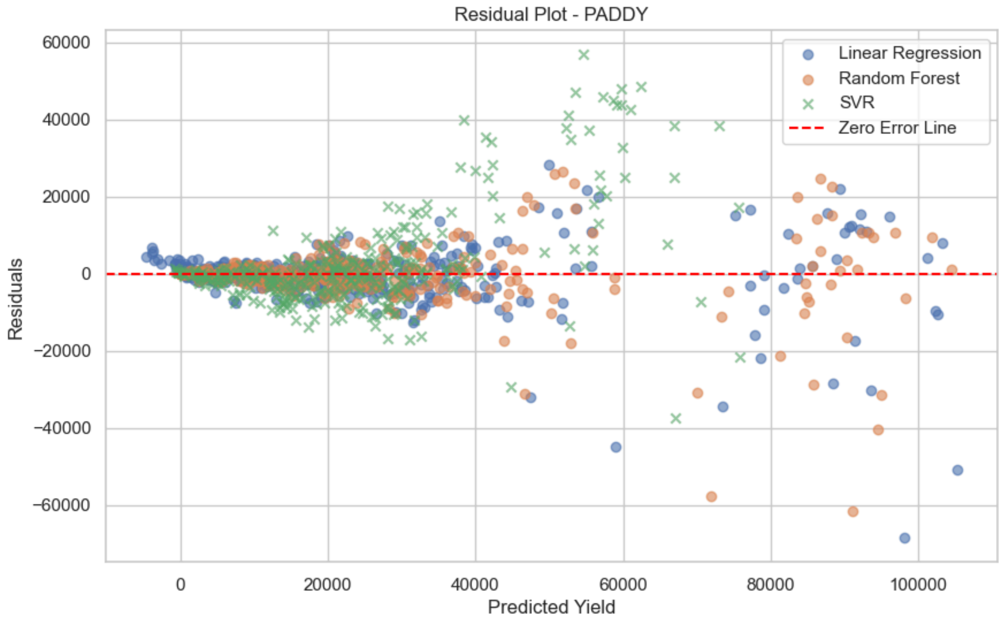


Figure 31. *Residual Plots for Paddy*

Note. This residual plot shows the differences between the actual and predicted crop yield for **Paddy** using the Random Forest model. The **X-axis** represents the predicted yield, while the **Y-axis** shows the residuals. The **red dashed line** indicates the perfect prediction line (residual = 0). Points above the line indicate over-predictions, and points below indicate under-predictions. Ideally, the points should be randomly scattered around the zero line, suggesting a good model fit.

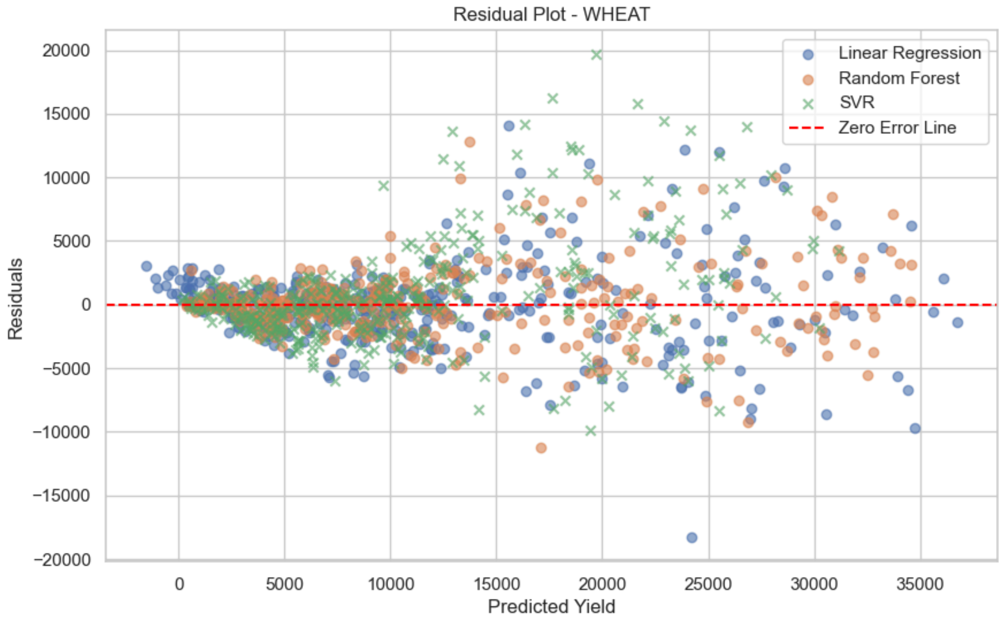


Figure 32. *Residual Plots for Wheat*

Note. This residual plot shows the differences between the actual and predicted crop yield for **Wheat** using the Random Forest model. The **X-axis** represents the predicted yield, while the **Y-axis** shows the residuals. The **red dashed line** indicates the perfect prediction line (residual = 0). Points above the line indicate over-predictions, and points below indicate under-predictions. Ideally, the points should be randomly scattered around the zero line, suggesting a good model fit.

4.5 Feature Importance

The Random Forest model's feature importance analysis revealed the top factors influencing yield per area for each crop. The degree of importance of these features is visualized in the following figures as pie charts. This allows us to understand the contribution of the independent factors towards the final crop yield.

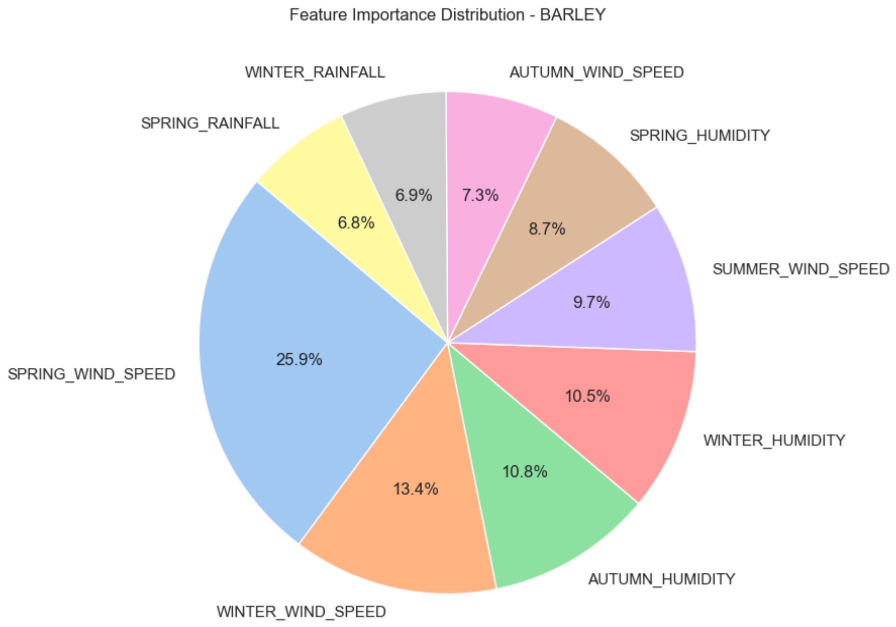


Figure 33. *Pie chart showing contribution of factors towards Barley yield*

Note. The pie chart displays the normalized importance of different features in predicting the crop yield for **Barley**. Each slice represents the contribution of a specific feature to the model's decision-making process, with the size of the slice indicating its relative importance. The chart helps to visualize which factors most influence the model's predictions, with features that contribute more taking up larger portions of the pie.

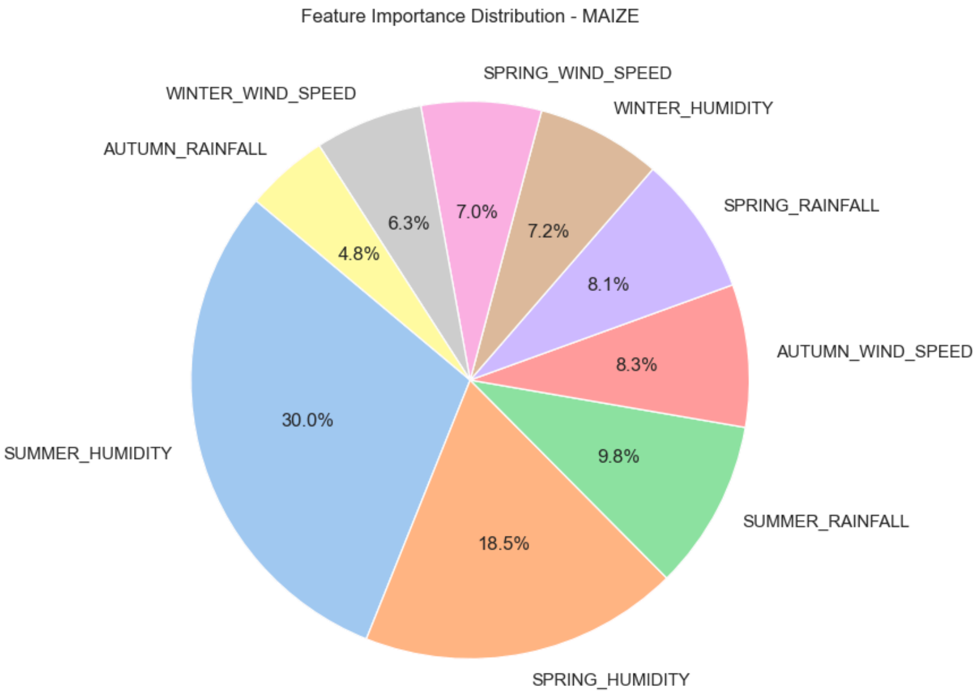


Figure 34. *Pie chart showing contribution of factors towards Maize yield*

Note. The pie chart displays the normalized importance of different features in predicting the crop yield for **Maize**. Each slice represents the contribution of a specific feature to the model’s decision-making process, with the size of the slice indicating its relative importance. The chart helps to visualize which factors most influence the model's predictions, with features that contribute more taking up larger portions of the pie.

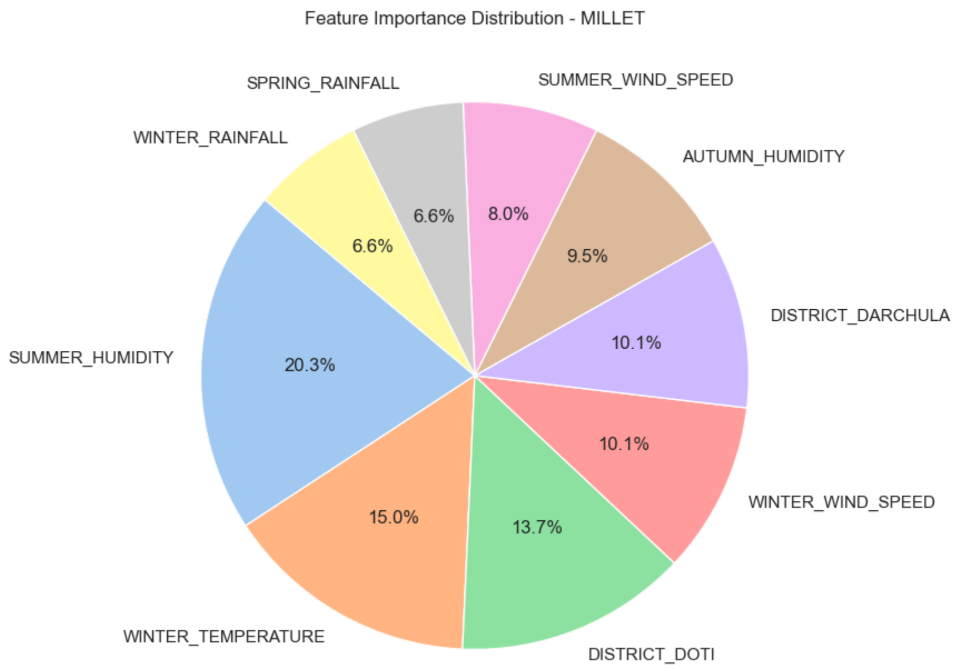


Figure 35. *Pie chart showing contribution of factors towards Millet yield*

Note. The pie chart displays the normalized importance of different features in predicting the crop yield for **Millet**. Each slice represents the contribution of a specific feature to the model's decision-making process, with the size of the slice indicating its relative importance. The chart helps to visualize which factors most influence the model's predictions, with features that contribute more taking up larger portions of the pie.

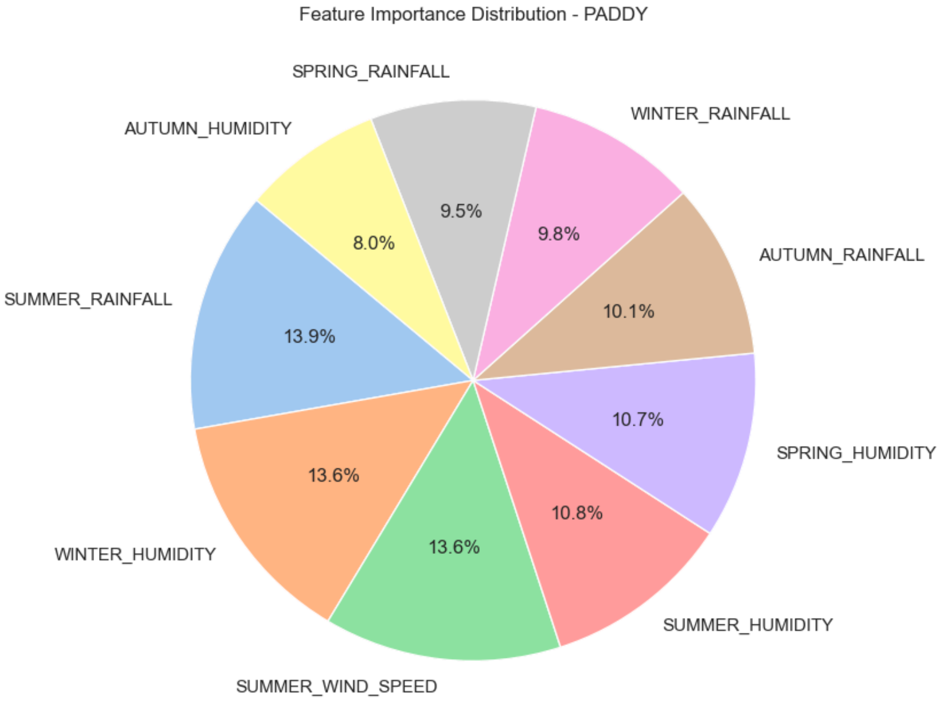


Figure 36. *Pie chart showing contribution of factors towards Paddy yield*

Note. The pie chart displays the normalized importance of different features in predicting the crop yield for **Paddy**. Each slice represents the contribution of a specific feature to the model’s decision-making process, with the size of the slice indicating its relative importance. The chart helps to visualize which factors most influence the model's predictions, with features that contribute more taking up larger portions of the pie.

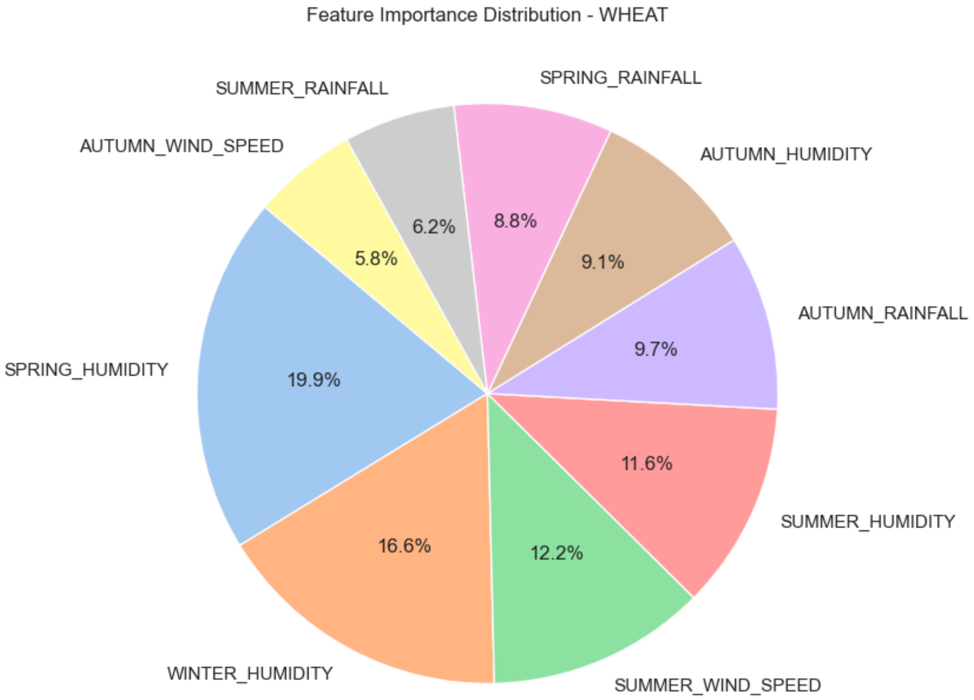


Figure 37. *Pie chart showing contribution of factors towards Wheat yield*

Note. The pie chart displays the normalized importance of different features in predicting the crop yield for **Wheat**. Each slice represents the contribution of a specific feature to the model’s decision-making process, with the size of the slice indicating its relative importance. The chart helps to visualize which factors most influence the model's predictions, with features that contribute more taking up larger portions of the pie.

CHAPTER 5

DISCUSSION

5.1 Model Performance Across Crops (From Tables)

5.1.1 Summary of Results

In this study, we evaluated three machine learning models—Random Forest, Support Vector Regression (SVR), and Simple Linear Regression—to predict crop yield based on weather and location data in Nepal. Our results in *tables 5 to 9*, indicated that Random Forest outperformed the other models, demonstrating superior accuracy in predicting crop yields across all tested cereal crops. This highlights the effectiveness of Random Forest for capturing non-linear relationships between weather variables and crop yield, which is essential in the context of agricultural forecasting (Araújo et al., 2023; Bondre & Mahagaonkar, 2019; Y. J. N. Kumar et al., 2020; Ruchira C. Mahore & Naresh G. Gadge, 2023; Saeed et al., 2017; Shahhosseini et al., 2020; Van Klompenburg et al., 2020).

The strong performance of Random Forest suggests that it can serve as a reliable tool for predicting crop yields in Nepal's diverse agricultural landscape. This has significant implications for enabling data-driven decision-making in agriculture, particularly in resource allocation, risk management, and policy formulation tailored to varying climatic and geographical conditions.

5.1.2 Key Observations

Random Forest consistently demonstrated superior performance across all crops, achieving the lowest MAE and RMSE values. This suggests that RF is better at capturing complex non-linear as well as linear relationships within the data and making more accurate predictions compared to LR and SVR (Bondre & Mahagaonkar, 2019; Y. J. N. Kumar et al., 2020; Morales & Villalobos, 2023). This strong consistency in the predicted values further proves RF to be the most viable machine learning algorithm to use.

While RF excelled in terms of prediction accuracy, SVR showed mixed results. It achieved the highest R^2 value for Barley, indicating a good fit to the data for that crop. However, SVR exhibited significantly higher MAE and RMSE values for Maize and Paddy, suggesting poorer predictive accuracy in general. The presence of outliers might have affected the performance of SVR while using the sensitive Linear kernel as reported by Zhang and O'Donnell (2020), and Haque et al. (2020) as well.

Linear Regression consistently showed the highest MAE and RMSE values across all crops, indicating lower accuracy compared to RF and SVR. However, it achieved relatively good R^2 values, suggesting that it can still capture a considerable portion of the variance in the data (Dongarra et al., 2011). It was also seen outperforming SVR in some instances, this might have occurred because that particular dataset was inherently linear in nature.

5.1.3 Insights

Going through the model performance summaries in tables 1 to 5, the dataset looks to be mostly linear with random bursts of complex relationships. It also contains extreme outlier that adversely affect model performance, specially for SVR, the presence of outliers skewed the predicted results by a significant amount (Üstün et al., 2007; Zhang & O'Donnell, 2020). The LR model was however not as affected by the outliers as the SVR model. This further reinforced that LR is excellent at capturing the simple baseline linear relation between the variables (Dongarra et al., 2011).

5.2 Predicted Vs Actual Yields (From Scatter Plots)

All through *figures 23 to 27*, we have the predicted yield in the y-axis and actual yield in the x-axis. We can also find a red dotted line along the center of the plot that denotes the perfect fit of the data that's coming from the model.

5.2.1 Overall Trends

The scatter plots for crops such as Maize, Paddy, and Wheat reveal distinct patterns for the three models. Random Forest shows a strong clustering of points along the diagonal (perfect fit line), indicating accurate predictions across a wide range of actual yield values. Support Vector Regression (SVR) demonstrates moderate accuracy but tends to spread more around the diagonal, especially at extreme yield values. Linear Regression exhibits significant deviation, highlighting its limited ability to model complex relationships in the dataset.

5.2.2 Performance across Barley and Millet

For crops like Barley and Millet, the scatter plots are overlapping and mostly indistinguishable from each other. This is further corroborated by the performance metrics from *tables 5,6 and 7* where either all of them performed good or all of them performed poorly. These trends that we see suggest the following:

- The dataset for these crops lack sufficient variability related information, and hence they might have trouble predicting values for extreme datasets (Ismail & Yusof, 2022; Q. Wang et al., 2017).
- All three models have captured the trends of the dataset in similar fashion for these two crops, this is likely due to the lack of external factors in the dataset like pest damage and control, and fertilizer usage.

5.2.3 Performance across Maize, Paddy and Wheat

The scatter plots for Maize, Paddy, and Wheat demonstrate clear distinctions in model performance, emphasizing the effectiveness of Random Forest in comparison to SVR and Linear Regression.

Random Forest shows a strong clustering of points along the diagonal (perfect fit line), indicating its superior ability to predict yields accurately. For these crops, the model captures the variations in yield effectively, even for mid-to-high yield values, where other models tend to falter. This can be attributed to Random Forest’s capability to handle complex, non-linear relationships between weather variables and crop yields (Schonlau & Zou, 2020).

SVR performs reasonably well, with predictions falling closer to the diagonal than Linear Regression but not as tightly clustered as Random Forest. Deviations become more pronounced at extreme yield values, where SVR struggles to capture the outliers. This limitation may stem from its reliance on kernel transformations that, while powerful, are less adaptable to highly heterogeneous data (Üstün et al., 2007; Zhang & O’Donnell, 2020).

Linear Regression exhibits significant scatter in the plots, particularly at higher actual yield values. This confirms its limited applicability for crops with non-linear yield patterns, such as Maize, Paddy, and Wheat. The consistent deviations indicate the model's inability to capture complex interactions between factors such as rainfall, temperature, and sunlight, as it is simply a linear equation trying its best to fit the data trend (Dongarra et al., 2011).

For these crops, the patterns in scatterplots reinforce the numerical results from performance metrics. The tighter clustering observed with Random Forest highlights its reliability, while SVR’s moderate deviations reflect its balanced but less robust predictions. Linear Regression’s widespread errors underscore the importance of using non-linear models for these crops.

This analysis highlights Random Forest as the most suitable model for predicting yields of Maize, Paddy, and Wheat, with potential applications in agricultural policy and planning. Accurate predictions for these high-production crops could guide decisions on resource allocation and risk mitigation strategies.

5.2.4 Patterns in Predictions

A pattern that kept repeating itself was that all models tend to underpredict when actual yields are exceptionally high, particularly for Maize and Wheat. This indicates that either the dataset is lacking entries for high yield situations or the models may not fully account for optimal environmental or management conditions leading to unusually high production (Bondre & Mahagaonkar, 2019; Ismail & Yusof, 2022).

5.3 Inferences from Residual Plots

The residual plots in *Figures 28 to 32*, across all crop yield models exhibit a near-conical shape. Specifically, smaller yield predictions tend to show smaller errors, while larger yield predictions display more significant errors. This pattern suggests heteroscedasticity in the data, where the variance of residuals increases with predicted yield values (Muller & Stadtmuller, 1987).

Several outliers were observed, notably in the cases of maize, millet, and wheat. These outliers could indicate either anomalies in the data or limitations in the model’s ability to generalize for certain yield ranges or specific conditions (Van Klompenburg et al., 2020). Identifying and addressing these outliers is critical for improving model robustness.

The conical pattern in the residuals implies that the model's predictions are less accurate for higher yield values. According to Kuradusenge et al. (2023), Muller & Stadtmuller (1987), and Van Klompenburg et al.(2020), this may stem from:

- **Data Imbalance**: A lower representation of high-yield scenarios in the dataset could lead to less effective model training for such cases.
- **Model Limitations**: The models, particularly the linear regression and SVR models, might struggle to capture complex relationships at the higher end of yield values.

- **Feature Interactions**: Higher yield values may depend on non-linear or complex interactions between features that were not adequately modeled.

5.4 Contribution Of Factors To Crop Yield (From Pie Charts)

The Random Forest model identified distinct features for predicting crop yields across the five cereal crops: Barley, Maize, Millet, Paddy, and Wheat. The results from *Figure 33* to *Figure 37* provide critical insights into the climatic and regional variables driving yield variability, shedding light on targeted interventions for improving agricultural outcomes.

5.3.1 Key Climatic Factors

Humidity emerged as a consistently significant factor across all crops. For example, **summer humidity (30.05%)** was the most influential variable for Maize, while **spring humidity (19.88%)** had the highest normalized importance for Wheat. The importance of humidity suggests its critical role in regulating crop growth, particularly during sensitive growth periods.

Rainfall was another dominant variable, particularly during summer. **Summer rainfall** showed the highest normalized importance for **Paddy (13.90%)** and was a significant factor for **Maize (9.84%)** and **Millet (9.53%)**. This underscores the heavy reliance of these crops on consistent monsoon rainfall for optimal yields.

Wind speed, often underemphasized in agricultural studies, appeared as a critical factor. **Spring wind speed** had the highest importance for **Barley (25.92%)** and was a major influence for **Millet (8.59%)** and **Maize (6.95%)**. This suggests that wind impacts factors such as pollination and soil erosion, which indirectly affect yield.

5.3.2 Regional Insights and Crop-Specific Observations

For Millet, district-level variability dominated feature importance, with regions such as **Gulmi (22.73%)** and **Gorkha (13.66%)** standing out. This highlights the significance of location-specific variables such as microclimates, farming practices, and soil conditions.

Paddy and Wheat showed a more balanced distribution of the features, with humidity, wind speed, and rainfall contributing significantly across seasons. The reliance of Paddy on **summer rainfall** and **summer wind speed (13.58%)** aligns with its sensitivity to monsoon conditions. In contrast, Wheat's dependence on **spring humidity (19.88%)** reflects its vulnerability to seasonal moisture variations during germination and growth phases.

5.3.3 Implications for Policy and Management

The identified key factors emphasize the importance of season-specific climatic conditions and regional characteristics in determining crop yields. Policymakers could leverage this information to:

1. Implement targeted irrigation systems for crops like Barley and Maize during dry seasons.
2. Strengthen monsoon management strategies for Paddy and Millet to reduce yield variability caused by erratic rainfall.
3. Focus on wind-mitigating measures in regions where crops like Barley and Millet are particularly affected.

These findings provide a foundation for developing localized agricultural interventions and improving the robustness of crop yield predictions, ultimately aiding resource allocation and food security initiatives in Nepal.

CHAPTER 6

CONCLUSION

This study explored the efficacy of machine learning models, specifically Random Forest (RF), Support Vector Regression (SVR), and Simple Linear Regression (SLR), in predicting crop yields based on weather and regional data for Nepal's five major cereal crops. Among the models, RF demonstrated superior performance, capturing complex, non-linear relationships in the data while providing interpretable influencing factors.

The analysis highlighted significant climatic factors, such as seasonal humidity, rainfall, and wind speed, which vary in importance depending on the crop. Regional factors, such as district-specific variability, further underscored the need for localized agricultural strategies. These insights emphasize the critical role of microclimatic conditions and their alignment with crop-specific requirements.

The findings hold substantial implications for policymakers, suggesting targeted interventions like improved irrigation systems during dry seasons and enhanced monsoon management strategies. Additionally, the study supports the utility of machine learning as a tool for guiding data-driven decisions in agriculture, aiding resource optimization and food security in Nepal.

While the results are promising, future research should focus on integrating additional datasets, such as soil properties and farming practices, to improve predictive accuracy further. Moreover, the deployment of predictive systems at the farmer level could enhance real-time decision-making, ensuring sustainable agricultural productivity in the face of climate change.

CHAPTER 7

RECOMMENDATIONS

Based on the findings of this study, the following recommendations are proposed to enhance agricultural productivity and decision-making:

1. **Integration of Advanced Technologies**: Agricultural agencies should adopt predictive analytics platforms powered by Random Forest or similar machine learning models to anticipate yield variability. Such platforms could support real-time decision-making for farmers and resource planners.
2. **Policy and Resource Allocation**: Policymakers should prioritize investments in irrigation infrastructure to mitigate the adverse effects of variable rainfall, particularly in districts with consistently low precipitation. Similarly, seasonal wind and humidity patterns should guide the allocation of resources to regions most vulnerable to yield fluctuations.
3. **Enhanced Data Collection**: Incorporating additional datasets, such as soil characteristics, pest prevalence, and farmer practices, into predictive models can improve accuracy and reliability. Developing a centralized, open-access database for agricultural and climatic data in Nepal could further support research and development.
4. **Further Research**: Future studies should explore temporal changes in feature importance due to climate change, ensuring that models remain relevant and adaptable. Additionally, interdisciplinary collaborations with environmental scientists and agronomists could refine predictions and recommendations.

REFERENCES

Adamopoulos, T., & Restuccia, D. (2018). *Geography and Agricultural Productivity: Cross-Country Evidence from Micro Plot-Level Data* (No. w24532; p. w24532). National Bureau of Economic Research. https://doi.org/10.3386/w24532

Agriculture Statistics | Publication Category | Ministry of Agriculture and Livestock Development. (n.d.). Retrieved January 13, 2025, from https://moald.gov.np/publication-types/agriculture-statistics/

Al-Gaadi, K. A., Hassaballa, A. A., Tola, E., Kayad, A. G., Madugundu, R., Alblewi, B., & Assiri, F. (2016). Prediction of Potato Crop Yield Using Precision Agriculture Techniques. *PLOS ONE*, *11*(9), e0162219. https://doi.org/10.1371/journal.pone.0162219

Ansarifar, J., Wang, L., & Archontoulis, S. V. (2021). An interaction regression model for crop yield prediction. *Scientific Reports*, *11*(1), 17754. https://doi.org/10.1038/s41598-021-97221-7

Araújo, S. O., Peres, R. S., Ramalho, J. C., Lidon, F., & Barata, J. (2023). Machine Learning Applications in Agriculture: Current Trends, Challenges, and Future Perspectives. *Agronomy*, *13*(12), Article 12. https://doi.org/10.3390/agronomy13122976

Archana, S., & Kumar, P. S. (2023). A Survey on Deep Learning Based Crop Yield Prediction. *Nature Environment and Pollution Technology*, *22*(2), 579–592. https://doi.org/10.46488/NEPT.2023.v22i02.004

Aworka, R., Cedric, L. S., Adoni, W. Y. H., Zoueu, J. T., Mutombo, F. K., Kimpolo, C. L. M., Nahhal, T., & Krichen, M. (2022). Agricultural decision system based on advanced machine learning models for yield prediction: Case of East African countries. *Smart Agricultural Technology*, *2*, 100048. https://doi.org/10.1016/j.atech.2022.100048

Ayoub Shaikh, T., Rasool, T., & Rasheed Lone, F. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Computers and Electronics in Agriculture*, *198*, 107119. https://doi.org/10.1016/j.compag.2022.107119

Bondre, D. A., & Mahagaonkar, S. (2019). PREDICTION OF CROP YIELD AND FERTILIZER RECOMMENDATION USING MACHINE LEARNING ALGORITHMS. *International Journal of Engineering Applied Sciences and Technology*, *04*(05), 371–376. https://doi.org/10.33564/IJEAST.2019.v04i05.055

C. Yang, C. L. Peterson, G. J. Shropshire, & T. Otawa. (1998). SPATIAL VARIABILITY OF FIELD TOPOGRAPHY ANDWHEAT YIELD IN THE PALOUSE REGION OF THE PACIFIC NORTHWEST. *Transactions of the ASAE*, *41*(1), 17–27. https://doi.org/10.13031/2013.17147

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., & Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, *274*, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010

Cantelaube, P., & Terres, J.-M. (2005). Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 476. https://doi.org/10.3402/tellusa.v57i3.14669

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003. https://doi.org/10.1088/1748-9326/aae159

Crop Yield Prediction and Fertilizer Recommendation. (2020). *International Journal for Research in Engineering Application & Management*, 135–138. https://doi.org/10.35291/2454-9150.2020.0452

Dahal, N., Ghimire, S., & Poudel, R. (2022). A Study on Dynamics of Major Cereal Crop Production in Nepal. *International Journal of Social Sciences and Management*, *9*(1), 13–18. https://doi.org/10.3126/ijssm.v9i1.42716

District Wise Climate Data for Nepal—District Wise Monthly Climate Data for Nepal—Open Data Nepal. (n.d.). Retrieved January 13, 2025, from https://opendatanepal.com/dataset/district-wise-daily-climate-data-for-nepal/resource/39cd1c13-2051-4d4c-9ea3-ba8106833166

Dongarra, J., Luszczek, P., Feautrier, P., Zee, F. G., Chan, E., Geijn, R. A., Bjornson, R., Dongarra, J., Luszczek, P., Philippe, B., Sameh, A., Philippe, B., Sameh, A., Dongarra, J., Luszczek, P., Steele, G. L., Gustafson, J. L., Dongarra, J., Luszczek, P., … Wismüller, R. (2011). Linear Regression. In D. Padua (Ed.), *Encyclopedia of Parallel Computing* (pp. 1033–1033). Springer US.

https://doi.org/10.1007/978-0-387-09766-4_2228

Elavarasan, D., & Vincent, P. M. D. (2020). Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications. *IEEE Access*, *8*, 86886–86901. https://doi.org/10.1109/ACCESS.2020.2992480

Fan, W., Chong, C., Xiaoling, G., Hua, Y., & Juyun, W. (2015). Prediction of Crop Yield Using Big Data. *2015 8th International Symposium on Computational Intelligence and Design (ISCID)*, 255–260. https://doi.org/10.1109/ISCID.2015.191

Fernandez-Beltran, R., Baidar, T., Kang, J., & Pla, F. (2021). Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. *Remote Sensing*, *13*(7), 1391. https://doi.org/10.3390/rs13071391

Ghimire, P. (2019). A Review of Studies on Climate Change in Nepal. *The Geographic Base*, *6*, 11–20. https://doi.org/10.3126/tgb.v6i0.26163

Government of Nepal, M. of F. (2024). *Economic Survey 2022/23*. https://www.mof.gov.np/site/publication-detail/3344

Gyawali, P., & Khanal, S. (2021). Overview of Agriculture in Nepal: Issues and strategies to cope with. *Fundamental and Applied Agriculture*, *0*, 1. https://doi.org/10.5455/faa.89753

Haque, F. F., Abdelgawad, A., Yanambaka, V. P., & Yelamarthi, K. (2020). Crop Yield Analysis Using Machine Learning Algorithms. *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, 1–2. https://doi.org/10.1109/WF-IoT48130.2020.9221459

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error

(MAE): When to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Islam, M. D., Di, L., Qamer, F. M., Shrestha, S., Guo, L., Lin, L., Mayer, T. J., & Phalke, A. R. (2023). Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning. *Remote Sensing*, *15*(9), 2374. https://doi.org/10.3390/rs15092374

Ismail, N., & Yusof, U. K. (2022). RECENT TRENDS OF MACHINE LEARNING PREDICTIONS USING OPEN DATA: A SYSTEMATIC REVIEW. *Journal of Information and Communication Technology*, *21*. https://doi.org/10.32890/jict2022.21.3.3

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). Random Forests for Global and Regional Crop Yield Predictions. *PLOS ONE*, *11*(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

Kattel, R. R., & Nepal, M. (2022). Rainwater Harvesting and Rural Livelihoods in Nepal. In A. K. E. Haque, P. Mukhopadhyay, M. Nepal, & M. R. Shammin (Eds.), *Climate Change and Community Resilience* (pp. 159–173). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-0680-9_11

Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN Framework for Crop Yield Prediction. *Frontiers in Plant Science*, *10*, 1750. https://doi.org/10.3389/fpls.2019.01750

Khanal, N. R., Nepal, P., Zhang, Y., Nepal, G., Paudel, B., Liu, L., & Rai, R. (2020). Policy provisions for agricultural development in Nepal: A review. *Journal of Cleaner Production*, *261*, 121241. https://doi.org/10.1016/j.jclepro.2020.121241

Kim, N., Ha, K.-J., Park, N.-W., Cho, J., Hong, S., & Lee, Y.-W. (2019). A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015. *ISPRS International Journal of Geo-Information*, *8*(5), 240. https://doi.org/10.3390/ijgi8050240

Kong, X., Hou, R., Yang, G., & Ouyang, Z. (2023). Climate warming extends the effective growth period of winter wheat and increases grain protein content. *Agricultural and Forest Meteorology*, *336*, 109477. https://doi.org/10.1016/j.agrformet.2023.109477

Krupnik, T. J., Timsina, J., Devkota, K. P., Tripathi, B. P., Karki, T. B., Urfels, A., Gaihre, Y. K., Choudhary, D., Beshir, A. R., Pandey, V. P., Brown, B., Gartaula, H., Shahrin, S., & Ghimire, Y. N. (2021). Agronomic, socio-economic, and environmental challenges and opportunities in Nepal's cereal-based farming systems. In *Advances in Agronomy* (Vol. 170, pp. 155–287). Elsevier. https://doi.org/10.1016/bs.agron.2021.06.004

Kumar, A., Sarkar, S., & Pradhan, C. (2019). Recommendation System for Crop Identification and Pest Control Technique in Agriculture. *2019 International Conference on Communication and Signal Processing (ICCSP)*, 0185–0189. https://doi.org/10.1109/ICCSP.2019.8698099

Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020). Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector. *2020 5th International Conference on*

Communication and Electronics Systems (ICCES), 736–741. https://doi.org/10.1109/ICCES48766.2020.9137868

Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., Rukundo, P., Mtonga, K., Mukasine, A., Uwitonze, C., Ngabonziza, J., & Uwamahoro, A. (2023). Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize. *Agriculture*, *13*(1), 225. https://doi.org/10.3390/agriculture13010225

Kuwata, K., & Shibasaki, R. (2016). ESTIMATING CORN YIELD IN THE UNITED STATES WITH MODIS EVI AND MACHINE LEARNING METHODS. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *III–8*, 131–136. https://doi.org/10.5194/isprs-annals-III-8-131-2016

Lohitha Reddy, K., & Siva Kumar, A. P. (2023). Machine Learning Techniques for Weather based Crop Yield Prediction. *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, 1263–1268. https://doi.org/10.1109/ICAIS56108.2023.10073740

Malla, G. (2009). Climate Change and Its Impact on Nepalese Agriculture. *Journal of Agriculture and Environment*, *9*, 62–71. https://doi.org/10.3126/aej.v9i0.2119

Mallikarjuna Rao, G. S., Dangeti, S., & Amiripalli, S. S. (2022). An Efficient Modeling Based on XGBoost and SVM Algorithms to Predict Crop Yield. In S. Borah, S. K. Mishra, B. K. Mishra, V. E. Balas, & Z. Polkowski (Eds.), *Advances in Data Science and Management* (Vol. 86, pp. 565–574). Springer Nature Singapore.

https://doi.org/10.1007/978-981-16-5685-9_55

Morales, A., & Villalobos, F. J. (2023). Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, *14*, 1128388. https://doi.org/10.3389/fpls.2023.1128388

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Muller, H.-G., & Stadtmuller, U. (1987). Estimation of Heteroscedasticity in Regression Analysis. *The Annals of Statistics*, *15*(2). https://doi.org/10.1214/aos/1176350364

Nepal, S., Neupane, N., Koirala, S., Lautze, J., Shrestha, R. N., Bhatt, D., Shrestha, N., Adhikari, M., Kaini, S., Karki, S., Yangkhurung, J. R., Gnawali, K., Singh Pradhan, A. M., Timsina, K., Pradhananga, S., & Khadka, M. (2024). Integrated assessment of irrigation and agriculture management challenges in Nepal: An interdisciplinary perspective. *Heliyon*, *10*(9), e29407. https://doi.org/10.1016/j.heliyon.2024.e29407

Ngoune Liliane, T., & Shelton Charles, M. (2020). Factors Affecting Yield of Crops. In Amanullah (Ed.), *Agronomy—Climate Change and Food Security*. IntechOpen. https://doi.org/10.5772/intechopen.90672

Nyéki, A., & Neményi, M. (2022). Crop Yield Prediction in Precision Agriculture. *Agronomy*, *12*(10), 2460. https://doi.org/10.3390/agronomy12102460

Palanivel, K., & Surianarayanan, C. (2019). AN APPROACH FOR PREDICTION OF CROP YIELD USING MACHINE LEARNING

AND BIG DATA TECHNIQUES. *INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING AND TECHNOLOGY*, *10*(3). https://doi.org/10.34218/IJCET.10.3.2019.013

Paudel, D., Boogaard, H., De Wit, A., Van Der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., & Athanasiadis, I. N. (2022). Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, *276*, 108377. https://doi.org/10.1016/j.fcr.2021.108377

Petersen, L. K. (2018). Real-Time Prediction of Crop Yields From MODIS Relative Vegetation Health: A Continent-Wide Analysis of Africa. *Remote Sensing*, *10*(11), 1726. https://doi.org/10.3390/rs10111726

Pradeep, G., Rayen, T. D. V., Pushpalatha, A., & Rani, P. K. (2023). Effective Crop Yield Prediction Using Gradient Boosting To Improve Agricultural Outcomes. *2023 International Conference on Networking and Communications (ICNWC)*, 1–6. https://doi.org/10.1109/ICNWC57852.2023.10127269

Priyadharshini, K., Prabavathi, R., Devi, V. B., Subha, P., Saranya, S. M., & Kiruthika, K. (2022). An Enhanced Approach for Crop Yield Prediction System Using Linear Support Vector Machine Model. *2022 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, 1–5. https://doi.org/10.1109/IC3IOT53935.2022.9767994

Rale, N., Solanki, R., Bein, D., Andro-Vasko, J., & Bein, W. (2019). Prediction of Crop Cultivation. *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 0227–0232. https://doi.org/10.1109/CCWC.2019.8666445

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches With Special Emphasis on Palm Oil Yield Prediction. *IEEE Access*, *9*, 63406–63439. https://doi.org/10.1109/ACCESS.2021.3075159

Renju, R. S., Deepthi, P. S., & Chitra, M. T. (2022). A Review of Crop Yield Prediction Strategies based on Machine Learning and Deep Learning. *2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS)*, 1–6. https://doi.org/10.1109/IC3SIS54991.2022.9885325

Ruchira C. Mahore & Naresh G. Gadge. (2023). Design A Model for Crop Prediction And Analysis Using Machine Learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 90–95. https://doi.org/10.32628/CSEIT228681

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saeed, U., Dempewolf, J., Becker-Reshef, I., Khan, A., Ahmad, A., & Wajid, S. A. (2017). Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan. *International Journal of Remote Sensing*, *38*(17), 4831–4854. https://doi.org/10.1080/01431161.2017.1323282

Sakib, S., Ahmed, N., Kabir, A. J., & Ahmed, H. (2018). *An Overview of*

Convolutional Neural Network: Its Architecture and Applications. MATHEMATICS & COMPUTER SCIENCE. https://doi.org/10.20944/preprints201811.0546.v1

Sakshi Mundhe, Jitu Sanap, Pooja Jadhav, Vaishnavi Kalsadkar, & Prof. Chiranjit Das. (2023). Forecasting Crop Yield For Sustainable Agriculture. *International Journal of Advanced Research in Science, Communication and Technology*, 29–34. https://doi.org/10.48175/IJARSCT-14205

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 3–29. https://doi.org/10.1177/1536867X20909688

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, *284*, 107886. https://doi.org/10.1016/j.agrformet.2019.107886

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4). https://doi.org/10.1214/15-AOS1321

Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting Corn Yield With Machine Learning Ensembles. *Frontiers in Plant Science*, *11*, 1120. https://doi.org/10.3389/fpls.2020.01120

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, *9*, 4843–4873.

https://doi.org/10.1109/ACCESS.2020.3048415

Sigdel, U. P., Pyakuryal, K. N., Devkota, D., & Ojha, G. P. (2022). Constraints on the use and adoption of information and communication technology (ICT) tools and farm machinery by paddy farmers in Nepal. *Journal of Agriculture and Forestry University*, 41–51. https://doi.org/10.3126/jafu.v5i1.48441

Tanabe, R., Matsui, T., & Tanaka, T. S. T. (2023). Winter wheat yield prediction using convolutional neural networks and UAV-based multispectral imagery. *Field Crops Research*, *291*, 108786. https://doi.org/10.1016/j.fcr.2022.108786

Tiwari, P., & Shukla, P. K. (2019). A Review on Various Features and Techniques of Crop Yield Prediction Using Geo-Spatial Data: *International Journal of Organizational and Collective Intelligence*, *9*(1), 37–50. https://doi.org/10.4018/IJOCI.2019010103

Üstün, B., Melssen, W. J., & Buydens, L. M. C. (2007). Visualisation and interpretation of Support Vector Regression models. *Analytica Chimica Acta*, *595*(1–2), 299–309. https://doi.org/10.1016/j.aca.2007.03.023

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Wan, C., Dang, P., Gao, L., Wang, J., Tao, J., Qin, X., Feng, B., & Gao, J. (2022). How Does the Environment Affect Wheat Yield and Protein Content Response to Drought? A Meta-Analysis. *Frontiers in Plant Science*, *13*, 896985. https://doi.org/10.3389/fpls.2022.896985

Wang, Q., Zhao, X., Huang, J., Feng, Y., Su, J., & Luo, Z. (2017). *Addressing Complexities of Machine Learning in Big Data: Principles, Trends and Challenges from Systematical Perspectives*. MATHEMATICS & COMPUTER SCIENCE. https://doi.org/10.20944/preprints201710.0076.v2

Wang, Y., Zhang, Z., Feng, L., Du, Q., & Runge, T. (2020). Combining Multi-Source Data and Machine Learning Approaches to Predict Winter Wheat Yield in the Conterminous United States. *Remote Sensing*, *12*(8), 1232. https://doi.org/10.3390/rs12081232

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.11172

Zhang, F., & O'Donnell, L. J. (2020). Support vector regression. In *Machine Learning* (pp. 123–140). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00007-9